基于语谱图融合的语音增强

Speech Enhancement Based on Spectrograms Fusion

学科专业：计算机科学与技术

作者姓名：史　昊

指导教师：王龙标　教授

答辩日期	2020 年 12 月 6 日		
答辩委员会	姓名	职称	工作单位
主席	陈彧	教授	天津理工大学聋人工学院
委员	于强	副教授	天津大学智能与计算学部
	贺瑞芳	教授	天津大学智能与计算学部

天津大学智能与计算学部

二〇二〇年十二月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **天津大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名: 史昊 签字日期: 2020 年 12 月 9 日

学位论文版权使用授权书

本学位论文作者完全了解 **天津大学** 有关保留、使用学位论文的规定。特授权 **天津大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

(保密的学位论文在解密后适用本授权说明)

学位论文作者签名: 史昊 导师签名: 王书标

签字日期: 2020 年 12 月 9 日 签字日期: 2020 年 12 月 9 日

摘　　要

在近些年，作为人机交互的一种重要方式，人们更加广泛地关注语音产品。语音产品的广泛应用，极大的解放了人们的双手，从而使日常生活变得更为方便。在较为干净的情况下，排除说话人自身的因素，语音应用已经能够取得很好的性能。例如，语音识别能够在干净的环境下，获得超过95%的准确率。但是，大量的噪声往往会在现实环境下存在，语音应用的性能也因此会受到极大的影响。从带噪语音信号中提取干净的语音信号的一种常用技术，被称为语音增强。

深度学习技术在近些年不断的发展，使得基于深度学习的语音增强方法吸引了越来越多研究者的关注。使用含有大量噪声种类和说话人的数据来训练一个深层神经网络，使得网络能够很好的处理非稳态噪声。深度学习语音增强的两种学习方式是直接映射和掩蔽。直接映射的方式是利用深层神经网络的强映射能力，利用深层神经网络直接得到谱图信息。而掩蔽的方式则是利用深层神经网络先得到一个掩膜，然后再用这个掩膜对带噪语谱图进行处理，最后得到增强的谱图信息。并且，通过直接映射和掩蔽得到的增强系统之间存在一定的互补性。

在这篇论文中，我们进一步利用这两个学习目标之间的互补性，设计了基于最小差别掩蔽的语谱图融合系统。利用最小差别掩蔽，提取不同语谱图中较好的部分。并将这些提取的部分重新融合成一张新的语谱图，从而提升语音增强的性能。在此基础上，我们使用了注意力机制更好的建模，并尝试了多种建模方式，来探究更有效的获得嵌入的方式。使用了增强语音的相位信息，来解决短时傅里叶逆变换时的不一致性。 REVERB数据集上的实验表明，语谱图和最小差别掩蔽之间具有很强的特征互补性。所提出的系统可以一致且显着地改善PESQ和SRMR，例如在所有真实数据中，平均SRMR增益为1.22。此外，我们的系统通过对信号干扰比率，信号失真比率，信号人造比率的感知评估，不断改进定量评估。

关键词： 语音增强，语谱图融合，最小差别掩蔽，深度学习

ABSTRACT

As an important way of human-computer interaction, people have gained widespread attention to speech applications in recent years. The wide application of speech applications has greatly liberated people's hands, thus making daily life more convenient. In a relatively clean situation, excluding the speaker's factors, speech applications have been able to achieve excellent performance. For example, in a clean environment, speech recognition can reach an accuracy rate of more than 95%. However, a lot of noises are existed in real life, and these noises will greatly affect the performance of voice applications. Speech enhancement is a common technique for extracting clean speech signals from noisy speech signals.

Deep learning technology has made big progress in recent years, deep learning-based speech enhancement methods have attracted more and more attention. The network can handle unstable noise well when using data containing a large number of noise types and speakers to train a deep neural network. Direct mapping and masking are two kinds of deep learning-based speech enhancement. The method of direct mapping is to use the strong mapping ability of the deep neural network and directly use the deep neural network to obtain the spectrogram information. The way of masking is to use a deep neural network to first obtain a mask, and then use this mask to process the noisy spectrogram, and finally obtain enhanced spectrogram information. Moreover, there is a certain complementarity between these two kinds of enhancement methods.

In this paper, we further utilize the complementarity between these two learning targets to design a spectrogram fusion system based on minimum difference masks. This system uses the minimum difference masks to extract the better parts of different spectrograms, and recombines these extracted parts into a spectrogram to improve the performance of speech enhancement.. On this basis, we used the attention mechanism for better modeling, and tried a variety of modeling methods to explore more effective ways to obtain embedding. The phase information of the enhanced speech is used to solve the inconsistency in the short-time inverse Fourier transform. The experiments on the REVERB challenge show that a strong feature complementarity between spectrograms and MDMs. Moreover, the proposed framework can consistently and significantly improve PESQ and SRMR, e.g., an average SRMR gain of 1.22 in all real data.

Besides, our system consistently improves the quantitative evaluation by the perceptual evaluation of speech quality, signal-to-distortion ratio, signal-to-interference ratio, and signal-to-artifact ratio.

KEY WORDS： Speech enhancement, spectrograms fusion, minimum difference masks, deep learning

目　录

第 1 章 绪论

本章首先对课题的研究背景，以及课题的研究意义进行介绍，即什么是语音增强、为什么要进行语音增强，以及对目前语音增强研究中存在的问题进行分析介绍，然后阐述近几年来语音增强的国内外研究现状，再根据现有研究中存在的问题提出本课题的研究内容以及本课题的贡献点，最后简述本文的整体组织结构。

1.1 研究背景及意义

近些年来，语音应用收到了很多的关注，例如苹果的Siri等，这极大的方便了人们的生活。人们可以直接用语音唤醒想要操作的设备，将语音作为人机交互的重要手段，发送指令，来操作机器，例如：在车载环境下，利用语音交互让系统播放歌曲或者调出导航；目前微信登陆时，可以利用声音锁来登陆等。但是，现实环境中往往存在大量的噪声[1]，而语音应用的性能则会受到这些噪声极大的影响[2,3]。在干净语音环境下，语音识别的性能可以达到95%以上的准确率，而在噪声存在时，语音识别的性能则会降低到40%甚至以下的准确率。因此，以什么样的方式来对带噪语音信号进行处理，进一步提升语音应用的性能是很重要的一项任务。

自IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 在2015发布 the 3rd CHiME Speech Separation and Recognition Challenge (CHiME-3) 挑战赛以来，语音增强任务显然得到了更多人的关注，许多更为困难的语音增强任务也被提出。 CHiME-3场景是在每天嘈杂的环境中使用的多麦克风平板电脑设备的语音识别。其中包含了四种不同的噪音设置：咖啡厅，路口，公共交通和步行区。由于CHiME-3的成功，于2016年发布的the 4th CHiME Speech Separation and Recognition Challenge (CHiME-4) 吸引了更多参赛者。如果说CHiME-3和CHiME-4已经是较难的任务，那么the 5th CHiME Speech Separation and Recognition Challenge (CHiME-5) 的提出，则将难度突然加大了很多。 CHiME-5的目标是在日常家庭环境中进行远距离麦克风对话语音识别的问题。语料库是从真实家庭中进行的二十次真实晚餐聚会中收集的。因为没有模拟数据，所以在真实数据上的使用，仍存在很多盲区，导致性能还不是很好。而2020年的the 6th CHiME Speech Separation and Recognition Challenge (CHiME-6) 仍是在CHiME-5基

础上，进一步探索提升在真实家庭场景下多说话人语音识别的性能的可能。

语音增强的目标即是从带噪语音信号中提取干净语音成分。语音增强可以分为两种[4]：传统的语音增强[5,6]和基于深度学习的语音增强[4]。传统语音增强方法往往会对语音信号进行数学建模的过程中，做一些假设，从而简化模型。传统语音增强算法在稳态噪声[7]时，往往能够取得很好的性能。但是这些语音增强系统在噪声处于非稳态时，他们的性能会急剧下降。此外，传统语音增强算法有时也会引入音乐噪声[8]，这也会影响语音应用的性能。深度学习在近些年有了特别大的发展，所以截止到目前为止，很多研究者将注意力投入到基于深度学习的语音增强研究中。深层神经网络具有十分强的非线性映射能力，利用这种强映射，不需要对语音信号假设，并且在处理非稳态噪声时也能取得更好的性能。

基于深度学习[9]的语音增强在近些年通过网络模型结构、学习方式等方面的研究，已经取得了很大的提升。但是，其学习目标可分为两类：基于直接映射作为学习目标的语音增强和基于掩蔽作为学习目标的语音增强[4]。虽然这两种学习目标在多目标学习的框架中，被证实出有一定的互补性[10]。但是，目前仍然缺少更多的研究，来验证并利用通过这两个学习目标训练的语音增强系统之间的互补性。

为了回答上面的问题，我们在本论文中提出了基于语谱图融合的语音增强系统。我们设计了最小差别掩蔽，并且利用最小差别掩蔽，提取多张语谱图当中较好的部分，并且将这些部分重新融合成一张新的语谱图。本方法进一步利用了基于直接映射和掩蔽的语音增强系统之间的互补性，可以为后续的研究做铺垫。因此，本文研究的基于语谱图融合的语音增强任务具有一定的研究价值和意义。

1.2 研究现状及存在的问题

本文的语音增强着眼于提升人耳听觉感受，使人们能更舒服的听带噪语音信号。虽然传统语音增强方法在处理稳态噪声[7]时，有很好的性能。但是，传统语音增强的性能当噪声处于非稳态时，这些系统的性能会收到极大影响。此外，利用传统语音增强方法处理带噪信号，有时则会引入音乐噪声[8]。近些年，深度学习理论被不断完善和发展，基于深度学习的语音增强方法[4]也受到了广泛的应用。深层神经网络极强具有很强的非线性映射能力，研究者利用这种映射能力，采用监督式学习[11]方式，对时域或者频域的语音特征进行处理，并还原成增强后的语音信号。

在基于深度学习的频域语音增强方法中。 Xu等[12]提出的框架，据我们了

解，是最早一批将深度学习技术应用到语音增强应用中的工作。首先，利用受限玻尔兹曼机（Restricted Boltzmann Machine, RBM）[13]将深层神经网络（Deep Neural Network, DNN）逐层预训练。然后，再利用整体调优的方式，对深层神经网络训练。相较于传统语音增强方法 log-minimum mean squared error （L-MMSE）[14]，此方法在多种噪声类型下都取得了较好的性能。基于上述工作，Xu et al.[15]发现特征输入到网络前，先采用拼帧处理，有益于语音增强任务。拼帧是指在预测当前帧时，引入当前帧的前几帧和后几帧，将这些帧作为一个整体，作为一个特征输入到神经网络中。除了对输入特征的改进外，使用较大的训练数据集训练神经网络能得到更好的增强性能。此外，在训练数据中加入多种类的噪声也有益于深度学习语音增强的鲁棒性[16]。

除了这些研究，很多研究者也对神经网络的结构进行了深入且全面的比较。 Hub等人在对猫的视觉皮层中细胞的研究基础上，提出卷积神经网络（Convolutional Neural Network, CNN）[17]。卷积神经网络是特殊设计的拟生物大脑皮层构。卷积层、池化层、激活函数、全连接层这四个部分组成了卷积神经网络。神经网络的复杂程度通过使用局部连接、权值共享和降采样这几项操作而被极大的降低。并且，通过引入多卷积核，可以让网络提取多个特征空间的特征。而循环神经网络（Recurrent Neural Network, RNN）[18]则通过引入时序的概念。尽管可以通过帧扩展将时间信息合并到DNN训练中，但由于没有明确地建模相邻帧之间的关系，因此在建模长期声学环境时仍然存在局限性。循环神经网络可以通过使用前一帧和当前帧之间的递归结构来捕获长期上下文信息并做出更好的预测来缓解此问题。利用不同种类网络各自的特点，将多种网络设计到一个模型中[19]，会使得语音增强获得更好的增强效果。

此外，一些设计的神经网络结构也能在语音增强任务上取得较好的效果。U-Net[20]结构在很多任务上被证明有效，同样可以在语音增强任务中取得很好的效果。 U-NET结构首先通过多次下采样，提供整个图像中的上下文语义信息。然后，经过级联操作从编码器直接传递到同高度解码器上的高分辨率信息，极大地为分割、提供更加精细的特征信息。生成对抗网络（Generative Adversarial Network, GAN）[21]通过使用生成和对抗的思想，使用鉴别器来判断生成器的效果，并通过引入鉴别器的某些信息，在一定程度上减轻了生成器可能带来的负面影响。生成对抗网络启发自博弈论中的二人零和博弈，整体网络包含一个生成模型（Generator, G）和一个判别模型（Discriminator, D）。生成模型生成输入带噪的语音特征，输出增强后的语音特征；判别模型是一个二分类器，判别输入是干净特征还是增强特征。当增强的特征分布越接近干净语音特征时，判别模型输出的概率越高，反之亦然。对抗生成模型的优化过程是一个“二元极小极大博弈（minimax two-player game）”问题[22]：在训练生成器时，先将判别器固

定，更新生成器模型的参数；而训练判别器时，则将生成器固定，只专注于训练判别器模型；生成器和判别器交替迭代训练，使得对方的错误最大化。此外，通过引入判别模型的损失，也可以缓解均方误差（Mean Squared Error, MSE）[23]作为损失函数的缺陷可能不适合人类听力的问题。

神经网络的学习方式对于语音增强的性能提升也是十分重要的。最小均方误差作为损失函数可能无法模拟人耳听觉模型。为了更适合人类的听力，一些语音增强系统还使用评估指标作为损失函数，例如，短时目标清晰度（Short-time objective intelligibility, STOI）[24,25]。通过计算网络预测的输出和标签之间的短时目标清晰度的误差，再将此误差以反向传播的方式训练深层神经网络，从而得到更适合人耳听觉感受的网络输出。还有些模型，会根据不同的信噪比（Speech-to-Noise Ratio, SNR）[26]训练不同的网络[27]。首先，网络会分别训练一个分类器和针对不同信噪比的多个语音增强模型。分类器的任务是区分当前需要增强语音的信噪比，而每一个信噪比都会有一个对应的语音增强模型。结合分类器来选择对应信噪比的语音增强模型，来降低网络模型泛化的负面影响。也有一部分工作针对低信噪比的情况，通过渐进式学习方式[28]逐步提高结果。基本思想是从小处着手，学习任务中较容易的方面或子任务，然后逐渐增加难度等级。特定于基于深层神经网络的语音增强，在深层神经网络训练中，从嘈杂语音到纯净语音的直接映射过程被分解为多个阶段，并且每个阶段都实现了信噪比增益。每个阶段的信噪比增益可以促进下一阶段的后续学习。简单堆叠多个深层神经网络[29]也显示了对语音增强任务有性能的提升。增强任务和其他任务的组合也可以彼此增强，例如：语音后验图（Phonetic Posteriorgrams, PPG）[30]与增强之间存在一定的相关性[31]，在结合训练时，可以提升两个任务的性能。

迁移学习[32]是另一个在语音任务中，很常用的一种训练方式。监督学习任务往往能够得到很好的性能，训练一个性能较好的深层神经网络需要大量的标注数据，具有大量标注好的数据对于监督学习任务时十分重要的。但是标注数据是一项枯燥无味且花费巨大的任务，而迁移学习的提出则是为了解决利用无标签数据的问题，也因此受到越来越多的关注。转移学习可以通过在大型数据库上进行训练并在小型数据库上进行微调来提高模型在更贴近于小型数据库上数据分布的性能。多目标学习（Multi-target Learning, MTL）[4]作为一种特殊的迁移学习[33]自从被提出后，就受到广泛的关注。通过共享神经网络中的很多参数，在同一个模型中，同时得到多种特殊输出，将多种任务同时处理。在利用多目标学习时，当多目标学习的任务是相关或存在一些互补性时，多目标学习训练的深层神经网络往往要优于单目标学习训练得到的深层神经网络。通过将梅尔频率倒谱系数（Mel Frequency Cepstral Coefficents, MFCC）[34]、掩蔽和语谱

图同时作为输入和输出，网络取得了相较于单独对语谱图输入输出系统更优的性能[10]。此外，利用多目标学习，语音增强任务也可以和其他语音任务，在一个模型中同时完成多种不同的任务，例如语音识别任务[35]。

近两年，随着端到端（End-to-end）[36]概念在语音增强上的不断发展，通过深层神经网络，在时域波形[37]上直接对带噪语音信号的时域波形进行处理，成为了现实。在许多文献中，时域语音增强基本上使用卷积神经网络或完全卷积神经网络（Fully Convolutional Neural Network, FCN）[38]作为网络结构。采用时域语音增强的好处是可以避免引入相位信息。经过长时间的研究，研究人员发现相位信息对于语音增强也很重要[39]。然而，由于没有固定的相位信息定律并且存在诸如相位缠绕之类的问题，因此难以处理相位信息。因此，在基于深度学习的早期频域语音增强中，仅处理幅度信息，然后以噪声相位重构波形[40]。考虑逆傅里叶变换（Inverse Short-time Fourier Transformation, ISTFT）时增强频谱图和噪声相位之间的不一致[39]。处理相位的最简单方法是迭代[41]。首先，利用增强后的幅度信息和噪声相位重建波形，然后提取重建波形的相位，然后使用重新提取的相位重建波形。通过重复迭代，可以改善效果。时域语音波形信号通过短时傅里叶变换（Short-time Fourier Transformation, STFT）[42]将信号以频域形式表示。语谱图是[43]频域语音增强系统的最常见特征。频谱图是一个三维结构：一个轴表示时间，第二个轴表示频率，第三个轴表示在特定时间特定频率的幅度。另外，由于提取的特征在T–F域中表示，因此需要相位信息[44]来重构波形，并且信号再次在时域信号中表示。虽然时域语音增强模型已经有了很大的提升，但是目前更多的工作还是集中于频域语音增强方法中。

尽管基于深度学习的语音增强发展已经有了长足的进步，但是目前学习目标只有两种：直接映射或者基于掩蔽的方法。直接映射是利用深层神经网络的强映射能力，利用网络直接预测得到增强后的语谱图[43]。而基于掩蔽的方法则是利用深层神经网络先预测得到掩蔽，然后利用预测的掩蔽对带噪语谱图进行噪声抑制。此外，基于掩蔽的方式和直接映射方法在不同情况下显示出不同的效果[18]，这显示出一些互补性。虽然通过多目标学习的方式可以利用此互补性，目前对于这两种学习目标获得的语音增强系统之间的互补性没有更多的研究。

1.3 主要工作及贡献

本课题主要研究基于语谱图融合的语音增强，进一步探索基于直接映射和基于掩蔽的语音增强得到的语谱图之间的互补性。主要包括基于多目标学习的语音增强系统、基于最小差别掩蔽的语谱图融合系统及基于注意力机制的语谱图融合系统三个部分。

（一）基于最小差别掩蔽的语谱图融合系统：我们设计了最小差别掩蔽（Minimum Difference Masks, MDM）来提取多个语谱图中较好的部分。然后基于此，设计了语谱图融合系统，在混响数据集上设计相关实验。构建语谱图融合系统存在两个挑战。首先是真实场景的非线性性质。尽管以前通过平均语谱图进行线性融合显示出良好的性能，但它无法通过简单的线性处理融合具有不同模式的各种系统。第二个是建立大规模的融合系统是不现实的。

1）在第一阶段，我们使用多目标学习并根据进行屏蔽和映射，以获取不同的学习目标频谱图，而不是构建具有大量资源的各种系统。对于非线性频谱图融合，我们设计了最小差别掩蔽来对T-F点进行分类，该箱最接近频谱图中的标签。

2）在第二阶段，使用深度神经网络对最小差别掩蔽进行估算，以获取不同频谱图的最佳部分。我们在第一阶段使用语谱图图，在第二阶段使用最小差别掩蔽将第一阶段得到的语谱图图重组为一个语谱图。

（二）基于注意力机制的语谱图融合系统：基于最小差别掩蔽的语谱图融合系统虽然在混响数据集上展现了很强的去混响性能，但仍存在很多问题。首先，使用最小差别掩蔽将频谱图的最佳部分融合到一个新的频谱图中，可能会破坏神经网络预测的频谱图的数据分布，从而导致频谱图不连续。其次，在融合过程中可以获得多个语谱图。尽管如此，它仍不使用这些语谱图来获取新的相位信息来代替原始的噪声相位，这很可能导致语谱图不一致。

在最小差别掩蔽的语谱图融合系统的基础上，我们提出了基于注意力机制的语谱图融合系统：

1）为了获得更好的频谱图连续性，我们在损失函数中添加了一个正则项。

2）为了减轻频谱图的不一致，我们使用线性融合波形的相位来重构最终波形，因为迭代信号重构可以产生更好的重新合成语音。

3）为了获得更好的神经网络建模能力，采用了注意机制。我们已经尝试了多种嵌入多个频谱图的方法作为注意力机制的输入。

1.4 论文结构

针对基于语谱图融合的语音增强研究，本文章从五个章节分别展开论述，具体的论文结构组织安排如下所示：

在本章中，我们首先介绍了本文的研究背景及意义，然后我们着重介绍了研究现状及目前存在的问题。最后，我们介绍本文的主要工作及本文的贡献。

在第二章中，我们介绍语音增强模型。我们首先介绍了语音增强的框架及特征提取，然后我们介绍基于深度学习的语音增强模型及在本文中会使用到的

评测指标。

在第三章中，我们首先对基于最小差别掩蔽的语谱图融合系统进行描述。然后我们分别给出最小差别掩蔽和基于语谱图融合的语音增强的详细模型介绍。最后，我们展示在REVERB数据集上的实验结果。并对实验结果进行分析和讨论。

在第四章中，我们完善基于最小差别掩蔽的语谱图融合系统，提出了基于注意力机制的语谱图融合系统。我们首先整体介绍此系统，然后介绍注意力机制，加入怎样的正则项来限制网络学习。此外，我们设计了多种网络嵌入方式，作为注意力机制的输入，来探究哪一种建模方式更有效。最后，我们用线性融合波形的相位信息替换带噪语谱图的相位信息来还原波形。实验的内容在本章也显示。

第五章首先对研究生以来的科研工作进行总结，然后对未来的研究方向进行了展望。

第 2 章　语音增强模型介绍

本章主要介绍频域语音增强的方法。首先，我们介绍语音增强的框架，并对常用的特征提取方法进行介绍。然后，我们介绍基于深度学习的语音增强方法。最后，给出本文中使用的评测指标。

2.1　语音增强框架

我们可以将麦克风接收到的语音信号建模为：

$$y(t) = s(t) + n(t), \tag{2-1}$$

其中y和s分别表示接收到的和干净的语音信号，而n表示加性噪声。 t表示的是时间点。 y首先采用短时傅立叶变换来表示频域中的时域信号，然后获得相位和语谱图。最常见的基于深度学习的语音增强系统，是使用深层神经网络，从带噪语音信号y获得增强的语谱图。通过使用逆短时傅立叶逆变换，从增强的语谱图和噪声相位重构s。

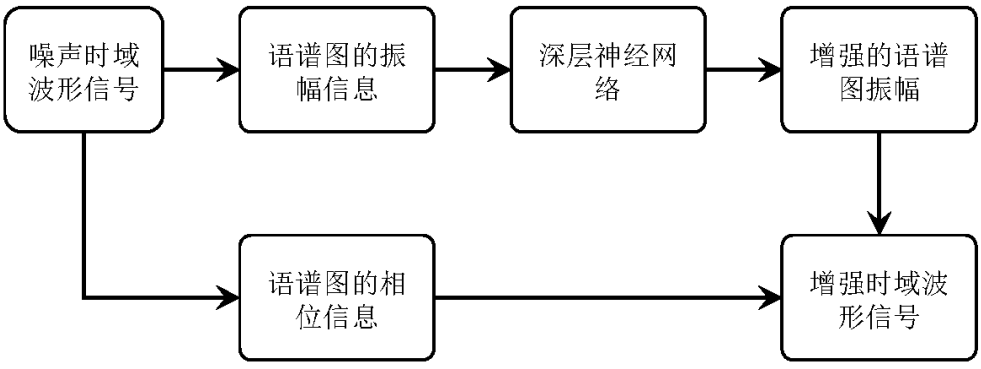


图 2-1　语音增强框架:利用深层神经网络，增强语谱图的振幅信息，然后将增强后的语谱图振幅信息和带噪语谱图的相位信息重构成增强时域波形。

2.2　传统的特征提取方法

虽然现在基于深度学习的语音增强已经十分流行，但是在这些系统中，仍然使用传统的语音特征提取方法提取到的特征。在本小节中，我们会介绍语谱图，语谱图是目前在语音增强中最常用的特征之一。

2.2.1 语谱图图像

语谱图是频域语音增强系统的最常见特征。频谱图是一个三维结构：一个轴表示时间，第二个轴表示频率，第三个轴表示在特定时间特定频率的幅度。图2-2是一个语谱图的示例。

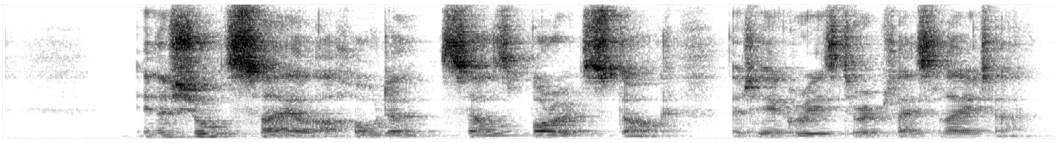


图 2-2　语谱图的示例：横轴表示的是时间，纵轴是频域，颜色深浅表示值的大小，值越大颜色越深。

根据加窗时窗的大小，可分为宽带语谱图和窄带语谱图。窄带语谱图具有非常好的频率分辨率，但是时间分辨率不理想。窄带语谱图，有精细的频率分辨率，共振峰的大致位置不能很好在图中体现出，即基波的变化特性不能被很好的展示。而宽带语谱图与窄带语谱图相反，宽带语谱图具有很好的时间分辨率，但是宽带语谱图具有较低的频率分辨率，声音的纹理特性无法很好的体现。随之而来的则是具有很好的时变特性，共振峰的大致位置可以很好的被分辨出，但是无法很好的分辨出谐波结构。

2.2.1.1 语谱图定义

语音信号的时域分析和频域分析是信号分析的两种重要处理分析方法，但是当单独使用这两种方法单独进行分析时，他们都有其局限性。具体表现在：语音信号的频率信息在时域分析的时候，没有直观的表示；而声音信号随时间的变化在频率信息分析的时候，则无法很好的表达。因此，为了充分分析信号的时间和频率信息，人们开始致力于研究信号的时频信息，将和时间序列相关联的傅里叶分析得出的图形叫做语谱图。语谱图是一种使用三维的方式来显示信号的频谱，其中纵轴代表的是频率信息，横轴表示的是时间信息，任意给定的频率信息和给定的时间信息对应的坐标点表示的是能量信息，通过灰度或颜色的深浅来表示，其代表的是能量的大小，颜色越深，表示该点的能量值越大，反之，则表示该点的能量值越小。

2.2.1.2 语谱图的提取

由于语音信号是一个非平稳态、时变的过程，随着时间的变化，其特性以及表征其本质特征的参数也是不断变化的。因此，在处理语音信号时，无法将一段长时间的语音信号当作平稳信号，以数字信号处理技术对这段语音进行处理和分析。但是，由于不同的语音信号是有声门的激励脉冲通过声道产生的，而

声道，即人的口腔肌肉运动相对于语音的频率来说是很缓慢的。在一个“短时间”的范围内，声道的形状是不变的，而这段语音信号我们一般认为其是稳态的、不随时间变化而改变的，一般这个“短时间”处于10毫秒到30毫秒之间，此外，男人和女人的时间是有差别的。由于假设语音信号在短时间内具有平稳的特性，所以人们常常认为在短时间内语音信号是一个准稳态的过程。因此，在提取语谱图时，首先对语音信号进行分帧处理，帧长一般选择10到30毫秒之间。为了使帧与帧之间能够保持相对的平滑，并且突出每一帧中的一部分信号，保持其连续不间断性，每一帧与每一帧之间设置有一定的重叠，即帧移。以帧移时间上的差别，每一帧提取得到的特征可以组成一个序列。帧移与帧长的比值通常设置为 0 – 1/2。通过可以移动的窗函数对语音信号进行加权的方法，叫做分帧操作。也就是用某一个窗函数$w(t)$乘以$s(t)$，从而形成加窗的语音或声音信号，即$s_w(t) = s(t) * w(t)$。窗函数一般选取汉明窗，其公式为：

$$fid = \begin{cases} 0.54 - 0.46cos\left[2\pi t/(T-1)\right], 0 \le n \le T-1 \\ 0, \qquad\qquad\qquad\qquad t = others \end{cases} \tag{2-2}$$

其中T表示帧长，$w(t)$表示窗函数。

语音信号或声音信号通过短时傅里叶变换将每一帧语音信号转到频域上进行处理。短时傅里叶变换的定义为：

$$X_t\left(e^{jw}\right) = \sum_{m=0}^{T-1} x(m)\, w(t-m)\, e^{-jwm} \tag{2-3}$$

有定义可以看出，窗选语音信号的标准傅里叶变换实际上就是短时傅里叶变换。窗函数$w(t-m)$是一个“滑动的”窗口，随t的变化而沿着语音信号序列$s(m)$滑动。最终每一帧特征参数组成特征参数的时间序列，时间信息以横轴表示，频率信息以纵轴表示，灰度表示能量大小信息的语谱图特征。

2.2.1.3 语谱图振幅和相位信息

语谱图包含振幅和相位两部分。我们以下面的公式表示对信号做短时傅里叶变换：

$$spec_y = STFT(y), \tag{2-4}$$

其中，$spec_y$是对y做短时傅里叶变换后的频域信号。语谱图振幅信息则可以表示为：

$$mag_y = |spec_y|^2, \tag{2-5}$$

而语谱图相位信息则可以表示为：

$$pha_y = phase(spec_y). \tag{2-6}$$

2.3 基于深度学习语音增强

基于深度学习的语音增强在近几年吸引了很多研究者的关注。目前，常见的学习目标可以分为两类：映射目标和掩蔽目标。学习目标对于基于深度学习的语音增强至关重要。在这一小节中，我们首先分别介绍直接映射的语音增强和基于掩蔽的语音增强。然后，我们介绍一种基于多目标学习的语音增强方法。

2.3.1 直接映射的语音增强

映射方法[12]使用神经网络直接获取增强的频谱图。映射目标的损失函数如下：

$$L_{MP} = \frac{1}{t * f} \sum_{t,f} (spc_{MP}(t,f) - spc_c(t,f))^2, \tag{2-7}$$

其中t和f分别表示时间和频率。 spc_{MP}是估计的映射频谱图， spc_c是干净的频谱图。 L_{MP} 是直接映射语音增强的损失函数。图2-3是直接映射的语音增强的框架。

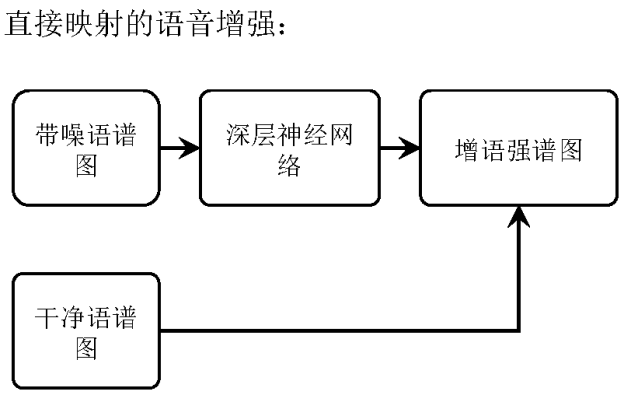


图 2-3 直接映射的语音增强的框架：利用深层神经网络的强映射能力，直接预测得到增强语谱图的振幅信息。训练时，需要利用干净语谱图的振幅信息与深层神经网络的输出计算损失量。

2.3.2 掩蔽的语音增强

基于掩蔽的语音增强包含两种，第一种是以掩蔽作为学习目标。这需要利用带噪语谱图和干净语谱图来计算一个理想的掩蔽。然后将此掩蔽作为学习目标，来监督训练。

另一种方式，信号近似（SA）[45]是屏蔽目标。它训练了一个比率掩码估计器，该比率估计器可以使干净语音的频谱幅度与估计语音的频谱幅度之间的差

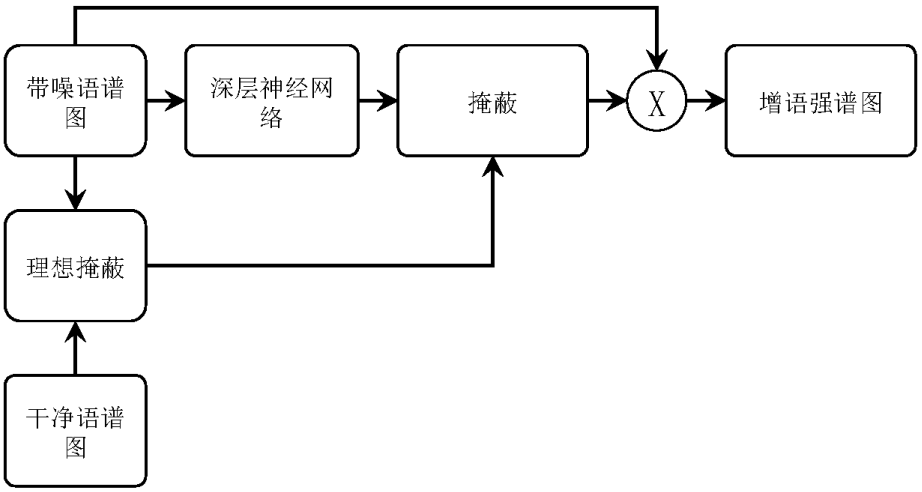
异最小。 SA的损失函数如下：

$$L_{SA} = \frac{1}{t*f}\sum_{t,f}(spc_{SA}(t,f) - spc_c(t,f))^2 = \frac{1}{t*f}\sum_{t,f}(spc_n(t,f)*mask(t,f) - spc_c(t,f))^2, \tag{2-8}$$

其中spc_n表示噪声频谱图， spc_{SA}是通过SA获得的频谱图，并且$mask$表示估计的掩码。在利用深层神经网络估计到掩蔽之后，我们利用此掩蔽来降噪。图2-4是基于掩蔽的语音增强的框架。

基于掩蔽的语音增强：

① 以掩蔽作为学习目标



② 以频谱作为学习目标（本文选用的方法）

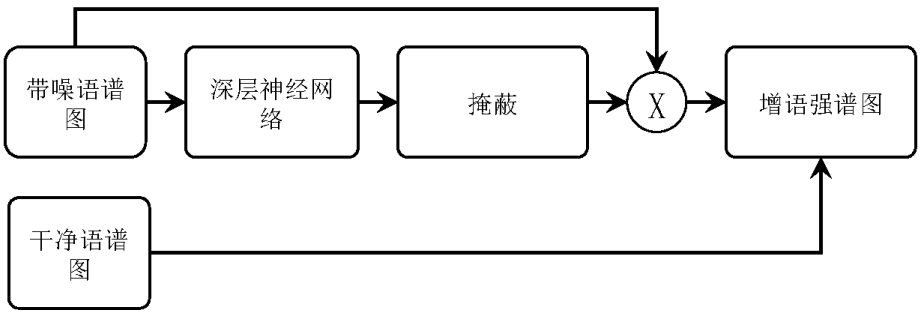


图 2-4　基于掩蔽的语音增强的框架，利用深层神经网络预测得到掩蔽值，可以分为以掩蔽作为学习目标和以谱图作为学习目标两类：以掩蔽作为学习目标时，需要先利用干净语谱图、带噪语谱图的振幅信息计算掩蔽值，然后利用这个计算好的掩蔽值作为标签，来训练深层神经网络；而以频谱作为学习目标时，需要将网络的输出首先与带噪语谱图相乘得到增强后的语谱图振幅信息，再将此振幅信息，与干净语谱图的振幅信息计算损失，训练网络。

2.3.3　多目标学习的语音增强

多目标学习[10,18]的想法是在一个模型中学习不同的目标。在多目标学习的

语音增强中，我们利用一个模型同时学到直接映射和基于掩蔽的语音增强系统。

$$L_{MTL} = L_{MP} + \alpha L_{SA}, \tag{2-9}$$

其中α是对应于神经网络对偶输出的两个MSE项的权重系数。这样，我们可以获得两个频谱图：直接映射的语谱图和基于掩蔽的语谱图。图2-5是基于掩蔽的语音增强的框架。

基于多目标学习的语音增强：

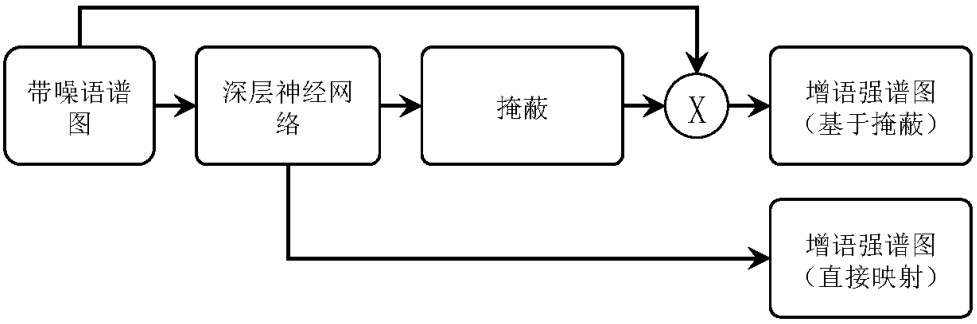


图 2-5　基于多目标学习的语音增强的框架利用深层神经网络，同时进行直接映射和掩蔽，可以得到基于映射的语谱图振幅信息和基于掩蔽的语谱图振幅信息。

2.4　评测指标

我们在本文中使用的评测指标有以下几个：

(-) 语音质量的感知评估（The perceptual evaluation of speech quality, PESQ）[46]。是ITU-T（国际电信联盟电信标准化部）在P.862 建议书中提供的客观MOS（Mean Opinion Scores）值评价方法。PESQ 引入了认知模型来描述参考语音信号与失真语音信号在听觉变换过程中产生的干扰差，通过模拟不对称和对称语音信号不同部分的不同加权改进了客观评估分值与MOS分值的相关性。

(-) 语音混响调制能量比（The speech-to-reverberation modulation energy ratio, SRMR）[47]。将混响语音作为输入，计算8 通道调制滤波器中前4 频带和后4 频带的调制能量比，将该比值作为可懂度客观预测分数。

(-) 信号失真比率（Signal-to-distortion ratio, SDR）[48]。是为了评估语音处理算法性能的一种客观度量指标，已公开在BSS_eval[48]工具箱中

(-) 信号干扰比率（Signal-to-interference ratio, SIR）[48]。

(-) 信号人造比率（Signal-to-artifact ratio, SAR）[48]。

2.5 本章小结

在本章中，我们首先介绍了频域语音增强的框架。然后着重介绍了语谱图是什么及其提取的方式。然后介绍了基于深度学习的语音增强方法，我们分别介绍了直接映射的语音增强、掩蔽的语音增强和多目标学习的语音增强。最后，我们介绍了本文使用的评测指标。

第 3 章　基于最小差别掩蔽的语谱图融合系统

语谱图融合是合并互补语音增强系统的有效方法。通过对多个语谱图求平均值，先前的线性语谱图融合显示出出色的性能。但是，具有不同功能的各种系统无法应用此简单方法。在这项研究中，我们设计了最小差别掩蔽（MDM），以根据距标签的最接近距离对频谱图中的时频点分档进行分类。然后，我们提出了一种用于语音增强的两阶段非线性语谱图融合系统。首先，我们进行基于多目标学习的语音增强前端模型，以同时获取语谱图。然后，估计最小差别掩蔽会采用不同语谱图的最佳部分。我们在第一阶段使用语谱图，在第二阶段使用最小差别掩蔽重组时频点。我们利用the REVERB challenge 数据集的实验表明，语谱图和最小差别掩蔽之间具有很强的特征互补性。此外，建议的框架可以持续且显着改善PESQ和SRMR，无论是真实数据还是模拟数据，例如，所有模拟数据的平均PESQ增益为0.1，所有真实数据的平均SRMR增益为1.22。

在本章中，我们首先对基于最小差别掩蔽的语谱图融合系统进行概述。然后，介绍最小差别掩蔽的方法。紧接着，我们介绍基于语谱图融合的语音增强。最后，我们对实验结果进行分析。

3.1　方法概述

语谱图融合系统旨在将多张互补的语谱图融合成一张新的语谱图。基于最小差别掩蔽的语谱图融合系统则是通过预测最小差别掩蔽，利用最小差别掩蔽分别提取相应语谱图中较好的部分，然后将这些部分重新融合成一张新的语谱图。

为此，我们设计了用于语音增强的非线性频谱图融合系统。现在，许多系统都根据最小均方误差准则进行训练，这激励了我们：如果将增强频谱图中接近干净频谱图的时频点区间融合回语谱图中，则可能对增强语谱图有所帮助：图3-1显示的是语谱图融合的方法概述。

在基于最小差别掩蔽的语谱图融合系统中：我们设计了最小差别掩蔽（Minimum Difference Masks, MDM）来提取多个语谱图中较好的部分。然后基于此，设计了语谱图融合系统，在混响数据集上设计相关实验。构建语谱图融合系统存在两个挑战。首先是真实场景的非线性性质。尽管以前通过平均语谱图进行线性融合显示出良好的性能，但它无法通过简单的线性处理融合具有不

同模式的各种系统。第二个是建立大规模的融合系统是不现实的。

1）在第一阶段，我们使用多目标学习并根据进行屏蔽和映射，以获取不同的学习目标频谱图，而不是构建具有大量资源的各种系统。对于非线性语谱图融合，我们设计了最小差别掩蔽来对T-F点进行分类，该箱最接近语谱图中的标签。

2）在第二阶段，使用深度神经网络对最小差别掩蔽进行估算，以获取不同语谱图的最佳部分。我们在第一阶段使用语谱图，在第二阶段使用最小差别掩蔽将第一阶段得到的语谱图重组为一个语谱图。

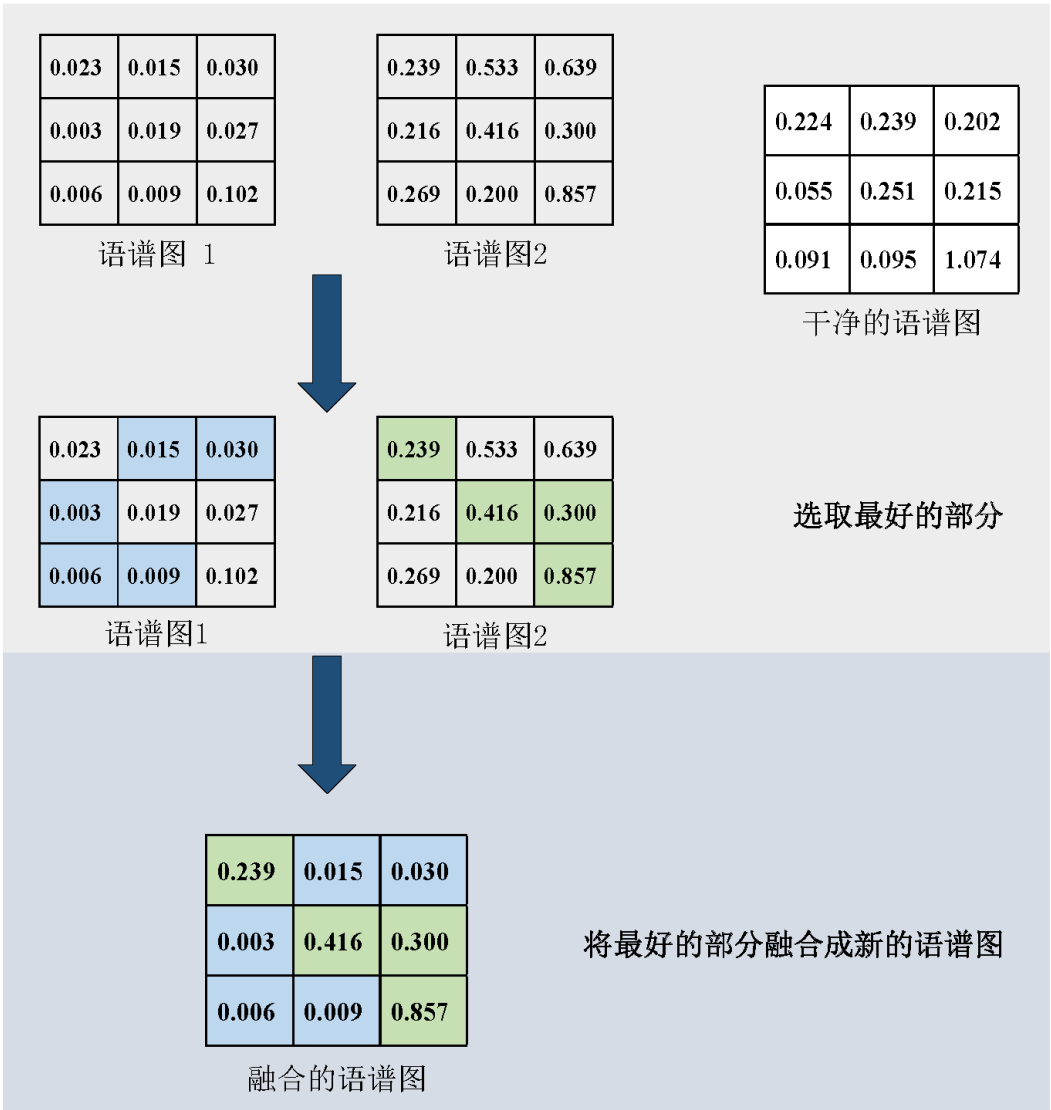


图 3-1　语谱图融合的方法概述：将多张语谱图中距离标签最近的时频点，重新组合成一个语谱图。

3.2 最小差别掩蔽

我们将每个增强型时频点与其相应标签之间的距离定义为 d_i：

$$d_i(t,f) = |spc_i(t,f) - spc_c(t,f)|, i \in \{MT-DM, MT-SA\} \tag{3-1}$$

其中 spc_i 表示多目标学习模型的增强频谱图。本研究中的 i 是 $MT-DM$ 或者 $MT-SA$。

最小差别掩蔽的标签定义为：

$$\widetilde{MDM}_i(t,f) = \begin{cases} 1, & i = \arg\min_i d_i(t,f) \\ 0, & otherwise \end{cases} \tag{3-2}$$

当 $d_i(t,f)$ 最小时，请将 $\widetilde{MDM}_i(t,f)$ 设置为值1，否则设置为0。

使用标签，最小差别掩蔽的估算可以视为一个监督问题。考虑到频谱图的连续性，最小差别掩蔽是测试中 (0, 1) 中的真实值。图3-2显示了计算最小差别掩蔽标签的过程。最小差别掩蔽的作用就是为了选取对应语谱图中相较干净语谱图距离较近的时频点。

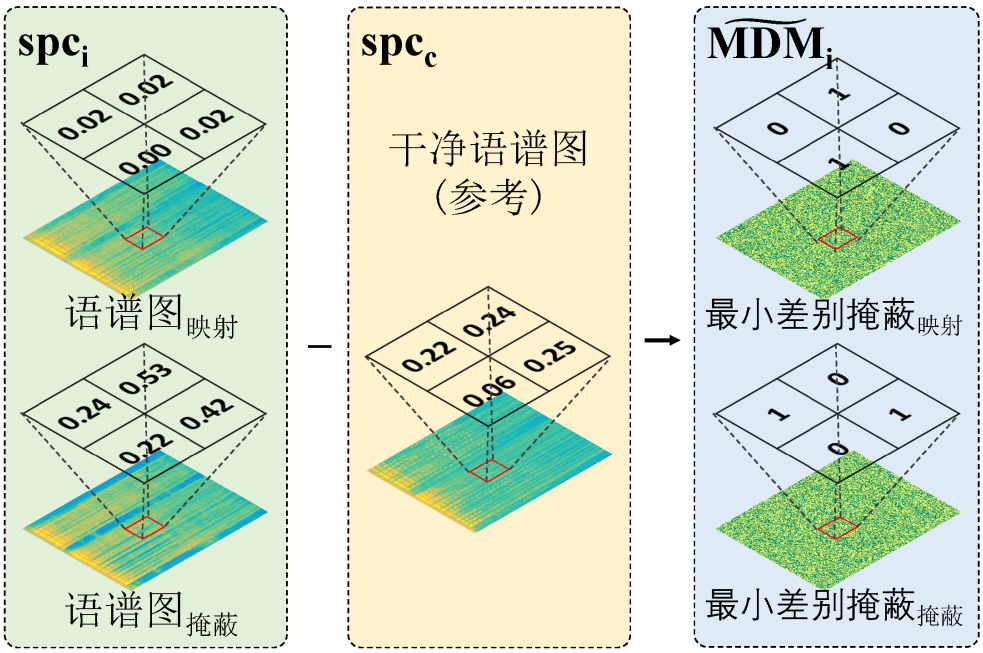


图 3-2 计算最小差别掩蔽标签的过程：spc_c 是干净的频谱图，spc_i 是从第一阶段开始的增强频谱图，$\widetilde{MDM}_i$ 是MDM的标签。将第一阶段多张语谱图与干净语谱图进行比较，选择距离标签最近的时频点。每一个语谱图对应一个最小差别掩蔽。

3.3 基于语谱图融合的语音增强

非线性频谱图融合包括两个阶段。在第一阶段，使用多目标的损失函数，进行基于多目标学习的语音增强前端模型，以获取不同目标的语谱图。然后训练

基于深层神经网络的后端模型来预测最小差别掩蔽。考虑到特征的互补性，进行了两个学习目标。仅估计最小差别掩蔽，并使用频谱图估计最小差别掩蔽：

$$L_{MDM-2O} = \sum_{i} \sum_{t,f} \left(MDM_i(t,f) - \widetilde{MDM}_i(t,f) \right)^2 \tag{3-3}$$

$$L_{MDM-4O} = L_{MDM-2O} + \alpha \left(L_{DM} + L_{SA} \right) \tag{3-4}$$

其中 $\widetilde{MDM}_i$ 表示MDM的标签，而 MDM_i 表示估计的MDM。我们将使用公式3-3训练的模型记为 **MDM-2O** ，而 **MDM-4O** 则表示用公式3-4训练的模型。

在测试阶段，进行非线性选择处理：

$$select_i(t,f) = MDM_i(t,f) * spc_i(t,f) \tag{3-5}$$

其中 $select_i$ 表示 spc_i 中的非线性选定部分。

最后，我们重新组合每个选定的部分以获得最终的增强频谱图：

$$spc_{fusion} = \sum_{i} select_i \tag{3-6}$$

其中 spc_{fusion} 表示最终的非线性融合语谱图。图3-3显示了第二阶段的训练和测试过程。

3.4 实验

在本节中，我们首先描述实验数据库。然后给出在本实验中深层神经网络的网络结构。最后，我们对实验结果进行分析。在本章中，我们在一个混响数据库上来测试。采用语音质量（PESQ）和语音与混响调制能量比（SRMR）性能的感知评估来对实验结果进行评估。

3.4.1 实验数据库

实验是在REVERB挑战任务上进行的。 REVERB挑战数据集包含模拟和真实话语；训练数据仅包括模拟记录。训练集共7861条模拟数据，验证集全部采用了模拟数据，而没有采用真实数据，这也意味着，我们在训练的过程中，未引入任何真实数据。测试数据的模拟和真实记录用于评估。语音信号被采样到16kHz。帧长和移位分别设置为512和256。输入和输出特征都是整个语音的语谱图的大小。

3.4.2 网络结构

所有网络均基于TensorFlow实施。在第一阶段，Bi-LSTM模型由257维输入层，两个隐藏层（每个层具有1024个节点）以及257维输出层组成，用于映射和掩蔽目标。在第二阶段，DNN模型由771维输入层，两个隐藏层（每层具

有1024个节点）和257维输出层组成，用于两个（MDM-2O）或（MDM-4O）四个输出。模型的参数是随机初始化的。采用一个验证集来控制学习率（初始为0.01），当两个连续迭代周期之间没有改善时，学习率将降低50%。此外，验证集中的性能决定了是否在一个时期内保存训练后的模型。每个时间反向传播（BPTT）或反向传播（BP）批次都包含八种发音。将第一阶段和第二阶段中用于多目标学习的 α 设置为1。

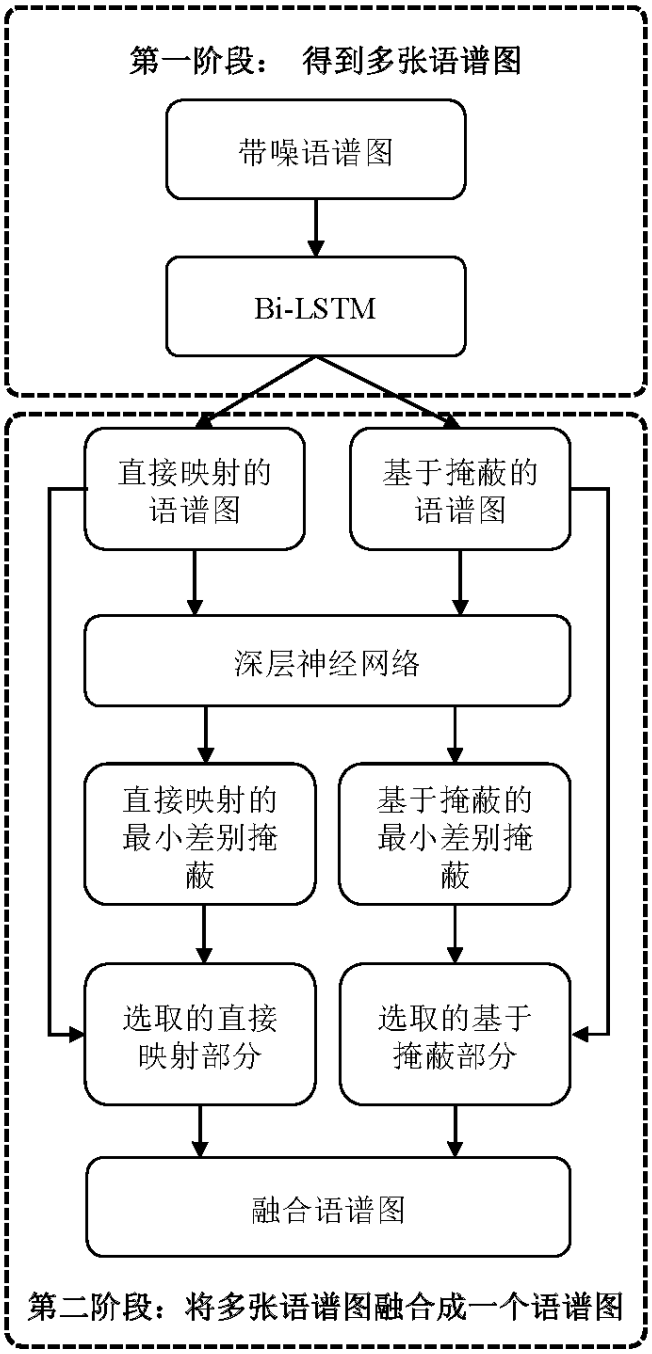


图 3-3 非线性频谱图融合系统：系统包含两阶段，在第一阶段，利用多目标训练方式训练双向长短时记忆网络，得到基于映射和基于掩蔽的两张语谱图；然后在第二阶段，将第一阶段得到的多张语谱图作为输入，利用深层神经网络，预测得到最小差别掩蔽，然后利用最小差别掩蔽，选取对应语谱图中较好的部分，最后重新融合成一个新的语谱图。

3.4.3 实验结果和讨论

表1显示了对模拟数据集的语音质量（PESQ）和语音与混响调制能量比（SRMR）性能的感知评估，表2显示了对实际数据集的SRMR性能。“Reverb”表示混响语音。“DM”和“SA”分别表示映射和掩蔽的方法。“MT-DM”和“MT-SA”表示使用等式的MTL方法的两个输出。“MT-LF”表示使用等式的线性融合谱图。所有基线模型均由257维输入层组成，两个隐藏层包含每层1024个节点。从结果中可以得出几个结论。

(1) 首先，对于PESQ指标，在远场中DM方法比SA方法产生更好的结果，而近场则相反。

(2) 第二，对于SRMR指标，SA方法始终优于DM方法。相反，DM方法产生的性能最差。

(3) 第三，多目标学习方法，MT-DM方法和MT-SA方法的每个输出是证明不同学习目标的互补性的最直接方法。因此，改进了PESQ和SRMR指标，过度使用了一个目标学习模型。

(4) 此外，多目标学习模型输出之一始终优于另一个。MT-SA方法比MT-DM方法具有更好的性能。

(5) 最后，尽管SRMR在远场中有一些退化，但线性语谱图融合有助于语音去混响。

表 3-1 模拟数据下的PESQ和SRMR结果：PESQ最大值是4.5，值越高表示性能越好；SRMR值越高表示性能越好。实验结果分别将数据集中近场数据（Near Room1、Near Room2和Near Room3）和远场数据（Far Room1、Far Room2和Far Room3）的三个房间数据做了平均。

模型	PESQ			SRMR		
	Far	Near	Avg.	Far	Near	Avg.
Reverb	2.15	2.59	2.37	3.43	3.94	3.68
DM	2.58	2.88	2.73	4.39	4.88	4.64
SA	2.54	2.93	2.74	4.48	4.92	4.70
MT-DM	2.56	2.90	2.73	4.42	4.92	4.67
MT-SA	2.60	3.01	2.81	4.64	4.97	4.80
MT-LF	2.64	3.02	2.83	4.58	4.99	4.78
MDM-2O(B)	2.56	2.92	2.74	4.38	4.54	4.46
MDM-2O	2.65	3.06	2.86	4.59	4.96	4.78
MDM-4O(B)	2.66	3.09	2.87	4.61	5.02	4.81
MDM-4O	**2.71**	**3.14**	**2.93**	**5.09**	**5.60**	**5.35**

“MDM-2O（B）”和“MDM-2O”的训练方法相同。但是，在融合中，“MDM-

2O（B）”在二进制掩码中将预测结果恢复为0-1值，而“MDM-2O”在实值掩蔽中使用预测概率。 “MDM-4O（B）”和“MDM-4O”和上面类似。表1和表2底部的结果表明，实值的掩蔽比二值掩蔽的效果更好，这表明软掩蔽比硬掩蔽更适合。此外，MDM-4O方法在所有PESQ和SRMR中都显示出其优越性。该结果表明，频谱图和最小差别掩蔽之间存在互补的活动功能。

表 3-2 真实数据的SRMR结果：SRMR值越高表示性能越好。实验结果分别将数据集中近场数据（Near Room1、Near Room2和Near Room3）和远场数据（Far Room1、Far Room2和Far Room3）的三个房间数据做了平均。

模型	SRMR		
	Far	Near	Avg.
Reverb	3.187	3.171	3.179
DM	3.291	2.926	3.109
SA	3.657	3.535	3.596
MT-DM	3.707	3.586	3.647
MT-SA	3.852	3.669	3.761
MT-LF	3.842	3.699	3.771
MDM-2O(B)	3.686	3.512	3.599
MDM-2O	3.931	3.767	3.849
MDM-4O(B)	3.956	3.815	3.885
MDM-4O	**5.055**	**4.927**	**4.991**

与MT-LF方法相比，除MDM-2O（B）方法外，大多数非线性频谱图融合方法均显示了出色的效果。使用二进制掩蔽融合频谱图可能会导致频谱图中的时变信息丢失，这可能是MDM-2O（B）方法性能不佳的原因之一。 MDM-2O和MDM-4O（B）在PESQ和SRMR中得到了更平滑的改进。相比之下，MDM-4O在真实和模拟数据上的PESQ和SRMR都有显着提高，例如，所有模拟数据的平均PESQ增益为0.1，所有真实数据的平均SRMR增益为1.22。 MDM-4O方法的成功启发了我们使用最小差别掩蔽作为辅助功能来预测未来工作中的语谱图或者掩蔽。

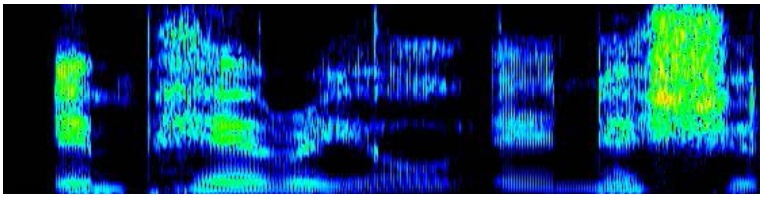
图3显示了幅度谱图。从图3可以得出一些观察结果。

(1) 首先，两种增强方法在减少混响和恢复掩埋在低频下的低频频谱方面均取得了极好的效果。

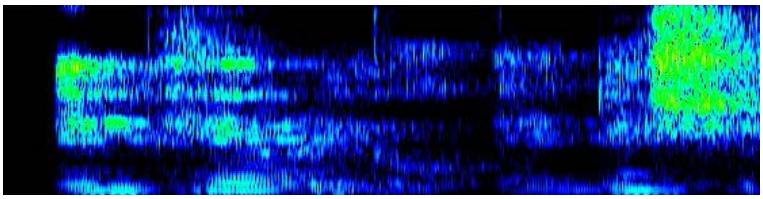
(2) 其次，干扰通常来自高频，MDM-4O方法具有出色的抑制高频干扰的能力。

考虑到我们在本研究中使用话语级特征来训练模型，作为比较实验，将探索具有帧扩展的帧级特征以在将来的工作中训练模型。而且，我们的方法具有

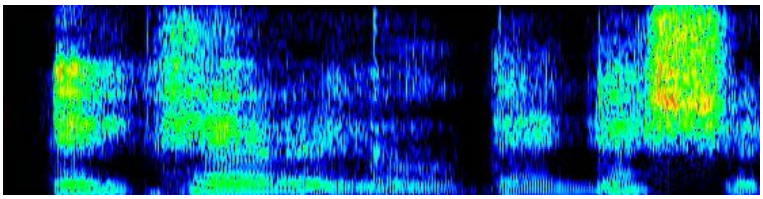
扩展到多系统融合的潜力。



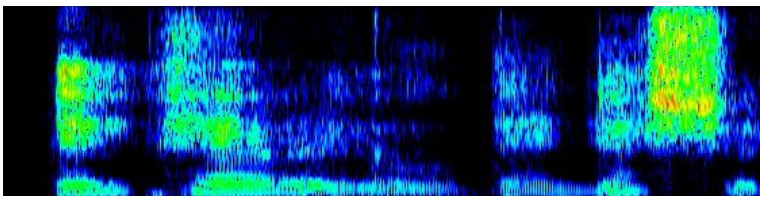
(a) Clean



(b) Reverberant



(c) MT-LF



(d) MDM-4O

图 3-4　语谱图的振幅信息：横轴表示时间，纵轴表示频率，颜色越深表示值越大，Clean表示干净的语谱图振幅信息，Reverberant表示带混响的语谱图振幅信息（未经过处理的带噪数据），MT-LF表示利用线性融合方式得到的增强语谱图振幅信息，MDM-4O表示利用本章节提出的基于最小差别掩蔽的语谱图融合系统得到的增强语谱图振幅信息。

3.5　本章小结

针对语音混响，我们提出了一种带有最小差分掩膜估计系统的非线性频谱图融合方法。基于最小均方准则，将增强语谱图中接近干净语谱图的时频点融合回语谱图中，则可能对增强语谱有所帮助。首先，进行了基于Bi-LSTM的多目标学习语音去混响前端，以获取不同的学习目标语谱图。利用多目标训练得到多个语谱图可以减少构建多个增强系统的资源浪费。然后，估计最小差别掩蔽将最接近标签的时频点分类。最后，我们使用第一阶段的频谱图和第二阶段的最小差别掩蔽来融合频谱图的最佳部分。当使用多目标学习时，我们观察到了频谱图和最小差别掩蔽（MDM）之间的主动特征互补。通过非线性频谱图融合，语音去混响主要提高了语音质量和语音混响调制能量比，例如，所有模拟

数据的平均PESQ增益为0.1，而实际数据的平均SRMR增益为1.22。在以后的研究中，我们将分析频谱图，并使用频谱图中的时变信息进行融合。此外，还将探索其他语音任务的特征融合，例如MFCC，以实现自动语音识别。

第 4 章　基于注意力机制的语谱图融合系统

在本章中，我们基于上一章基于最小差别掩蔽的语谱图融合系统，提出了基于注意力机制的语谱图融合系统。

我们提出了一种基于新的注意力机制的语谱图融合系统，该系统具有最小差别掩蔽（MDM）估计，用于加性噪声的语音增强。与以前使用全连接神经网络的系统相比，我们的系统利用了多头注意力机制。具体而言，我们（1）尝试将多种语谱图的嵌入方法作为关注机制的输入，这些方法可以提供频谱图中相邻帧之间的多尺度相关信息；（2）在损失函数中添加正则项以获得更好的语谱图连续性；（3）使用线性融合波形的相位来重构最终波形，这可以减少不一致的语谱图的影响。在MIR-1K数据集上进行的实验表明，我们的系统提高了语音质量的量化评估的水平。

我们首先概述整体算法，然后介绍注意力机制。并提出最小差别掩蔽的语谱图融合系统的缺点，并从引入损失函数的正则项、网络嵌入的设计和增强语音的相位信息几个方面来完善系统。最后，我们给出实验所用的数据集及实验分析。

4.1　方法概述

基于最小差别掩蔽的语谱图融合系统虽然在混响数据集上展现了很强的语音增强性能，但仍存在很多问题。首先，使用最小差别掩蔽将频谱图的最佳部分融合到一个新的频谱图中，可能会破坏神经网络预测的频谱图的数据分布，从而导致频谱图不连续。其次，在融合过程中可以获得多个语谱图。尽管如此，它仍不使用这些语谱图来获取新的相位信息来代替原始的噪声相位，这很可能导致语谱图不一致。

在最小差别掩蔽的语谱图融合系统的基础上，我们提出了基于注意力机制的语谱图融合系统：

1）为了获得更好的频谱图连续性，我们在损失函数中添加了一个正则项。

2）为了减轻频谱图的不一致，我们使用线性融合波形的相位来重构最终波形，因为迭代信号重构可以产生更好的重新合成语音。

3）为了获得更好的神经网络建模能力，采用了注意机制。我们已经尝试了多种嵌入多个频谱图的方法作为注意力机制的输入。

图4-1显示了基于注意力机制的语谱图融合系统。

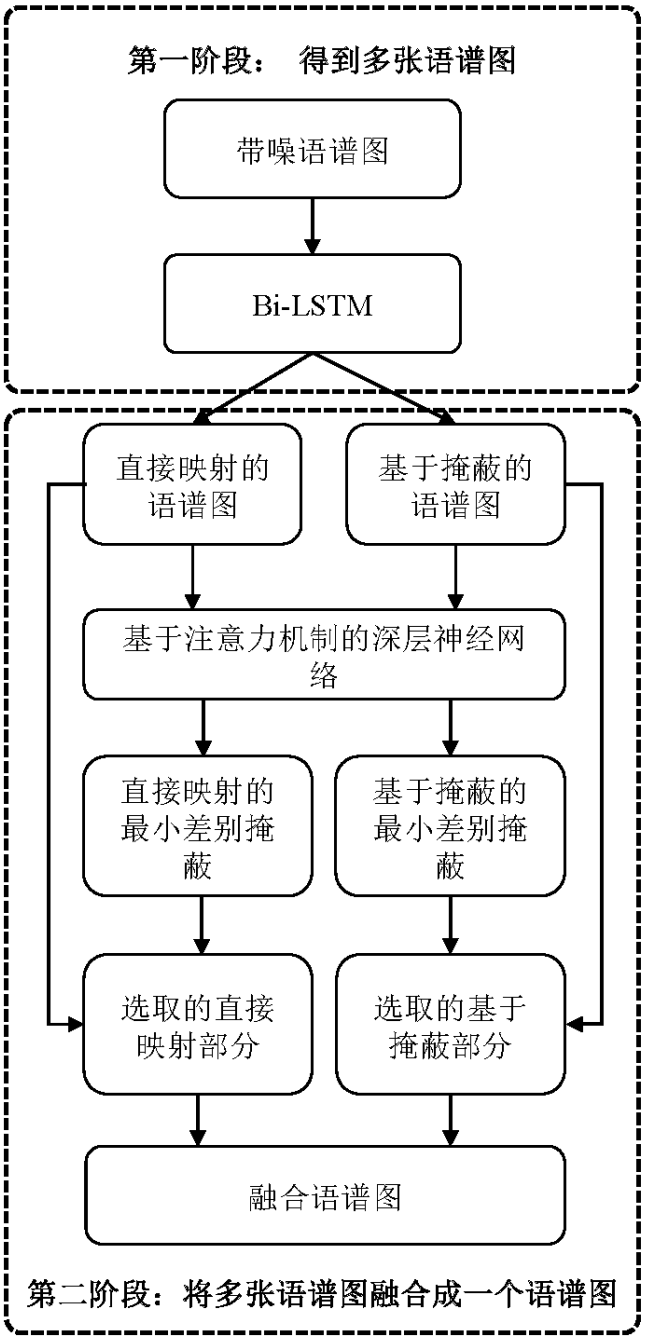


图 4-1　基于注意力机制的语谱图融合系统：系统包含两阶段，在第一阶段，利用多目标训练方式训练双向长短时记忆网络，得到基于映射和基于掩蔽的两张语谱图；然后在第二阶段，将第一阶段得到的多张语谱图作为输入，利用基于注意力机制的深层神经网络，预测得到最小差别掩蔽，然后利用最小差别掩蔽，选取对应语谱图中较好的部分，最后重新融合成一个新的语谱图。

4.2　注意力机制

注意机制可以描述为计算值的加权和，其中分配给每个值（V）的权重是通过查询（Q）与相应键（K）。我们对打包成矩阵Q的一组查询计算注意力机制。

键和值也打包在一起形成矩阵K和V。输出矩阵的计算如下：

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{4-1}$$

其中d_k是查询和键的维。

而多头注意力机制代替使用单一的关注机制，串联多个注意力机制允许模型共同关注来自不同位置的不同表示子空间的信息：

$$MultiHead(Q,K,V) = Concat(head_1,...,head_n)W^O \tag{4-2}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{4-3}$$

其中W_i^Q，W_i^K和W_i^V表示线性投影的参数矩阵。$head_i$表示第i注意力机制。多个注意力机制被连接并再次投影以给出最终值。图4-2是多头注意力机制。

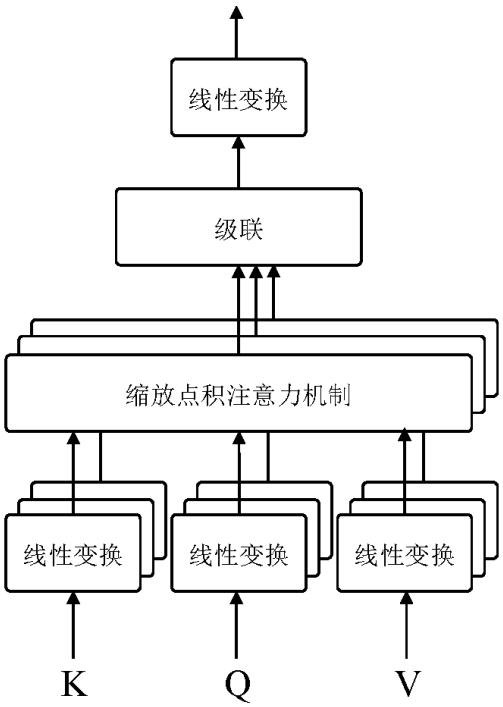


图 4-2　多头注意力机制：K表示键值（key），Q表示查询（Query），V表示值（Value），K，Q，V首先经过线性变换，然后将线性变换的结果使用缩放点积注意力机制，最后将注意力机制的结果级联，并再经过一次线性变换。

4.3　损失函数的正则项

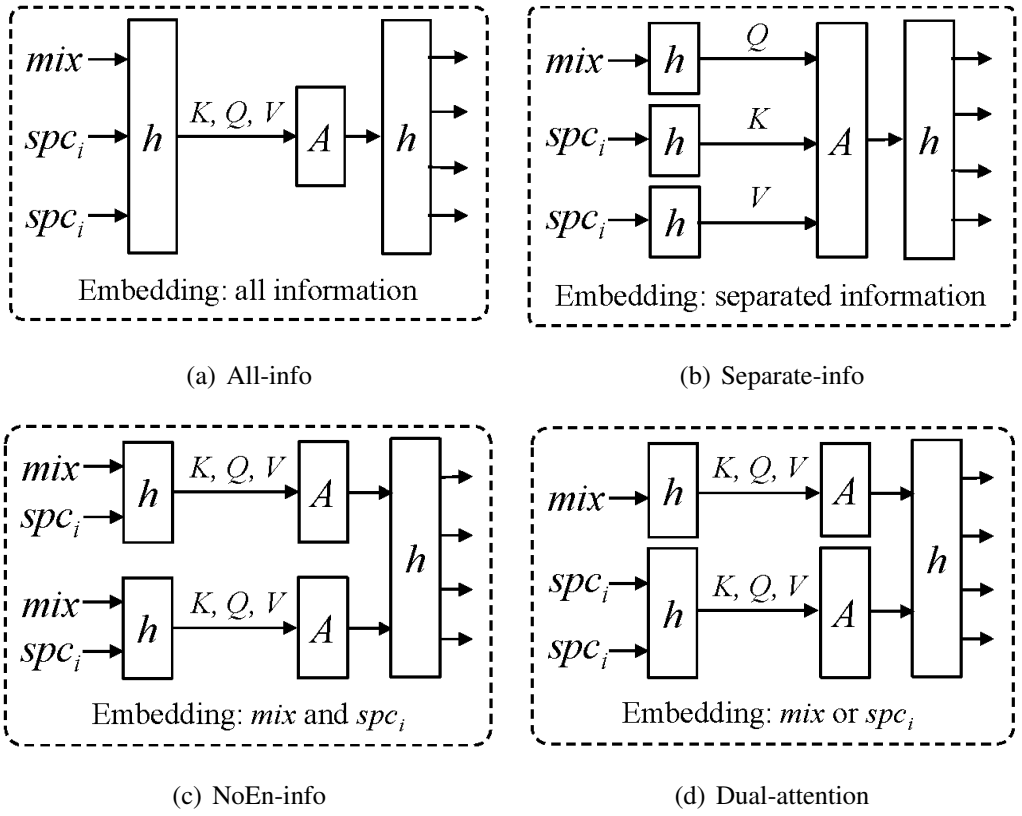
使用最小差别掩蔽将频谱图的最佳部分融合到一个新的频谱图中，可能会破坏神经网络预测的频谱图的数据分布，从而导致频谱图不连续。考虑到融合谱图的连续性，我们在学习过程中添加了一个正则项到损失函数中：

$$L_{MDM-tend} = L_{MDM} + \gamma\left(spc_f - spc_c\right)^2 \tag{4-4}$$

我们称使用此正则项的模型为**MDM-tend**。

4.4 网络嵌入

网络嵌入旨在将输入数据映射到潜在空间中，因此它是输入数据的另一种表示形式。此外，不同的输入数据或不同的网络嵌入方法可能会对嵌入效果产生重大影响。在本文中，我们使用隐藏层作为嵌入网络。



(a) All-info (b) Separate-info

(c) NoEn-info (d) Dual-attention

图 4-3 多种频谱图的嵌入方法作为关注机制的输入： *mix* 是带噪的语谱图；$spc_i (i \in (mapping, masking))$ 是增强的语谱图； *h* 是一个隐藏层， *A* 是注意力机制； *K*, *Q* and *V* 是注意力机制中的键，查询和值； (a) 所有信息作为嵌入 (All-info), (b) 信息作为单独嵌入 (Separate-info), (c) 带噪和增强的信息作为嵌入 (NoEn-info), (d) 带噪的信息和增强的信息分别建模 (Dual-attention)。

4.5 增强语音的相位信息

迭代信号重构可以产生更好的语音合成，因此我们使用线性融合波形和非线性融合频谱图的相位来重构最终的增强波形。线性融合语谱图（也称为集成谱）获得，可以通过 $spc_{mapping}$和 $spc_{masking}$ 这两个增强的语谱图获得。

$$spc_{LSF} = \left(spc_{mapping} + spc_{masking}\right) / 2 \tag{4-5}$$

在重构语音时，先使用线性融合谱图和带噪相位信息恢复成时域波形信号。

然后再利用此时域波形信号提取相位信息，以此相位信息来恢复最后增强的语音。

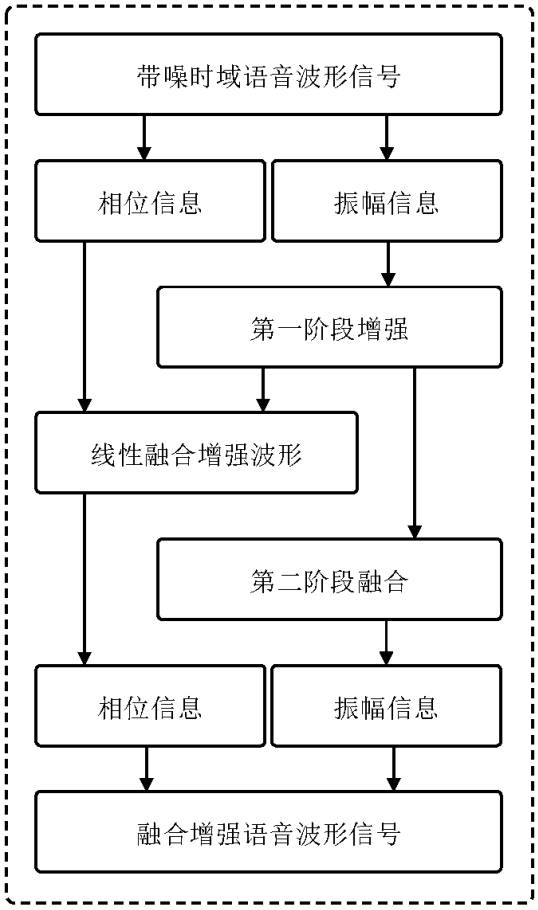


图 4-4 基于注意力机制的语谱图融合系统：首先将带噪时域语音波形信号特征提取，得到相位信息和振幅信息，利用深层神经网络，在第一阶段得到多张增强语谱图，并在第二阶段将第一阶段得到的多张语谱图融合得到融合后的增强语谱图振幅信息；相位信息则采用第一阶段得到的增强结果的线性融合语音信号的相位信息，将此相位信息和融合增强语谱图振幅信息重构成增强语音波形信号。

4.6 实验

在本节中，我们首先描述实验数据库。然后给出在本实验中深层神经网络的网络结构。最后，我们对实验结果进行分析。在本章中，我们在一个加性噪声数据库上来测试。我们对本节中很多名称做了归纳，如表4-1所示。

4.6.1 数据库

实验是在MIR-1K数据集上进行的。 MIR-1K数据集包含以16 kHz采样率和16位分辨率记录的1000个歌曲剪辑。这些剪辑包含混合音轨和音乐伴奏音轨，包括8位女性和11位男性的声音。我们选择了所有tammy的剪辑作为测试集，总共8个剪辑。随机选择十二个剪辑作为验证集，其余980个剪辑用作训练集。我

表 4-1　在本章中重要的变量名及其描述。

变量名	描述
Q	注意力机制的查询
K	注意力机制的键
V	注意力机制的值
$Attention(\cdot)$	注意力机制
$head_i$	第 i 个注意力机制
$MultiHead(\cdot)$	多头注意力机制
W	线性映射的参数矩阵
$d_i(t,f)$	增强语谱图时–频点和标签时–频点在 (t,f) 位置的距离
$\widetilde{MDM}_i(t,f)$	第 i 个最小差别掩蔽在 (t,f) 位置的标签
$MDM_i(t,f)$	第 i 个最小差别掩蔽在 (t,f) 位置的预测值
spc_i	第 i 个增强的语谱图
$select_i$	第 i 个增强语谱图中被选中的部分
spc_{fusion}	最终增强的语谱图

们合成了两条音轨以产生单声道混合唱歌语音数据，从而使信噪比等于0。

4.6.2　网络结构

所有网络均基于Tensorflow实施。模型的参数是随机初始化的。网络参数如表4-2所示。另外，我们使用验证集上的性能来决定是否在一个时期内保存经过训练的模型。因为映射和掩码都很重要，所以α被设置为1。在间隔（0，1）中β和γ之间的差异对结果影响很小，因此将它们分别设置为1和0.5。

表 4-2　语谱图融合系统的参数设置

设置	第一阶段	第二阶段
神经网络	Bi-LSTM	Attention + DNN
隐含层个数	2	1
每个隐含层的结点数	512	1024
输入特征维度	257	257 * 3
输出特征维度	257 * 2	257 * 4
学习率	0.01	0.01
训练轮数	30	30
批大小	8	8

4.6.3 实验结果和讨论

语音质量（PESQ），信号失真比（SDR），信号干扰比（SIR）和信号伪像比（SAR）的感知评估被用作评估指标。 “S-Masking”表示使用$L_{masking}$作为训练损失的屏蔽方法，而“S-Mapping”表示使用$L_{mapping}$作为训练损失的映射方法。“M-Mapping”和“M-Masking”表示使用等式的多目标学习方法的两个输出。 “M-LSF”表示线性融合的方法。 “uPIT-vocal”示使用uPIT训练的声音输出。

表 4-3 非线性融合方法的效果：MDM-tend表示利用加入正则项损失函数训练得到的系统；+phase表示利用MDM-tend增强的语谱图振幅信息和线性融合增强语音波形的相位信息重构的语音波形。

系统	SDR	SAR	SIR	PESQ
Mix signal	0.058	140.81	0.058	1.112
S-Masking	9.315	11.645	13.448	1.629
S-Mapping	9.324	11.496	13.743	1.914
M-Mapping	9.215	11.261	13.835	1.965
M-Masking	9.804	11.834	14.425	1.851
M-LSF[18]	9.770	11.934	14.161	2.090
uPIT-vocal[49]	9.751	11.902	14.141	1.854
MDM[50]	10.036	11.830	15.096	2.217
MDM-tend	10.063	11.848	15.142	2.212
+phase	**10.391**	**12.050**	**15.709**	**2.263**

从表4-3可以看出。映射和掩蔽方法对各个评测指标的影响不同。例如，对于PESQ，SDR和SIR，S-Mapping的结果要优于S-Masking，而SAR的结果则相反。多目标学习方法优于单一学习方法；即，M-Mapping和M-Masking始终显示出优越的度量。多目标学习模型的输出之一始终优于另一个。即M-Masking的效果比M-Mapping好。 uPIT-vocal方法显示出很强的声音分离能力，但是PESQ有所下降。 MIR-1K数据集上的实验与REVERB数据集非常相似。但是对于STL，S-Mapping方法产生的性能优于S-Masking。线性融合方法没有表现出良好的性能，而所提出的非线性融合方法对于语音增强仍然有效。对于PESQ测度，远场的S-Mapping效果优于远场的S-Masking，而近场则相反。对于SRMR度量，S-Masking始终优于S-Mapping，而S-Mapping产生了最差的性能。多目标学习方法的每个输出都是证明不同学习目标的互补性的最直接方法。因此，过度使用了一种单目标学习的模型，从而改善了PESQ和SRMR措施。多目标学习使多种学习目标之一可以学习更多； M-Masking的性能优于M-Mapping。尽管SRMR在远场中有一些退化，但线性融合有助于语音增强。

“MDM-tend”表示使用正则项训练的非线性融合方法。 “+ phase”表示重建波形时使用的线性融合波形的相位。在神经网络中添加正则项，更改频谱图的信息会得到更好的结果；例如，MDM趋向方法优于MDM方法。通过在线性融合方法的语音中提取相位，可以获得更好的相位。这样得出的平均PESQ增益为0.052，平均SDR增益为0.328，平均SAR增益为0.206，平均SIR增益为0.554。MDM方法在所有PESQ和SRMR中都显示出其优越性。该结果表明，频谱图和MDM之间具有很强的互补性。 MDM方法的成功启发了我们使用MDM作为辅助功能来预测未来工作中的频谱图或掩膜。

表 4-4　不同嵌入方式建模的结果：MDM-tend-All-info表示注意力机制的嵌入方式是将带噪语谱图振幅信息和增强后的两个语谱图振幅信息作为一个嵌入，输入到注意力机制中；MDM-tend-Separate-info表示将带噪语谱图振幅信息和增强后的两个语谱图振幅信息分别作为一个嵌入，输入到注意力机制中；MDM-tend-NoEn-info表示将带噪语谱图振幅信息分别和增强语谱图振幅信息作为嵌入，输入到注意力机制中；MDM-tend-Dual-attention表示将带噪语谱图振幅信息作为一个嵌入，将两个增强后的语谱图作为一个嵌入输入到注意力机制中。

系统	SDR	SAR	SIR
MDM-tend	10.391	12.050	**15.709**
MDM-tend-All-info	10.397	12.100	15.626
MDM-tend-Separate-info	**10.461**	**12.252**	15.491
MDM-tend-NoEn-info	10.417	12.132	15.613
MDM-tend-Dual-attention	10.397	12.152	15.503

表4-4可以得出一些结论。注意机制有助于对声谱图之间的关系进行建模，从而减少了语音的失真和干扰程度。此外，通过评估指标的波动，可以看出注意机制建模可以更好地减少加性噪声和音乐噪声。但是，伴奏的效果几乎没有降低。 MDM-tend-Separate-info显示出最佳性能；这意味着注意力机制可以从单个频谱图的嵌入中更好地学习信息。所有建模方法都有助于语音增强，从而验证了所提出方法的鲁棒性。没有一个系统可以在所有指标上获得一致的改进，这可能意味着关注机制以不同的建模方式获得了不同的信息。

更详细的不同的Q，K和V组合在表4-5中显示，以查看哪种注意力机制可以更好地对语谱图建模。从表4-4中，我们可以观察到以下几点：

1) 所有的K，Q和V组合均可提升语音增强的性能。这意味着注意力机制提供了更好的建模功能。

2) 当K和V采取相同的语谱图时，它们往往在某种程度上具有最佳效果。这可能是因为当K和V具有相同的数据分布时，它更有利于注意力机制的使用。

3) 对于相同的Q，掩蔽的效果要好于映射。基于掩蔽的语谱图优于基于映射的语谱图。也许更有效的语谱图将提供更有效的信息，并有助于对注意力机制

进行建模。

4) 最好将混合语谱图用作Q，这可能是因为增强语谱图中的某些信息丢失了。具有更完整信息的混合语谱图将有助于研究注意力机制。

表 4-5 不同Q，K和V的趋势注意力机制的最小差别掩蔽结果（+phase）；“mapping”表示来自MTL模型的映射频谱图，“masking”表示来自多目标学习模型的掩蔽频谱图，“average”表示线性融合频谱图，“mix”表示混合频谱图，*K*, *Q* and *V* 是注意力机制中的键，查询和值。

Q	K	V	SDR	SAR	SIR
mix	mapping	masking	10.421	12.230	15.414
mix	masking	mapping	10.445	12.238	15.470
mapping	mix	masking	10.402	12.162	15.500
mapping	masking	mix	10.385	12.163	15.443
masking	mix	mapping	10.397	12.142	15.530
masking	mapping	mix	10.405	12.181	15.465
mix	mapping	mapping	10.440	12.231	15.473
masking	mapping	mapping	10.393	12.135	15.532
average	mapping	mapping	10.355	12.129	15.425
mapping	mix	mix	10.382	12.133	15.500
masking	mix	mix	10.390	12.130	15.530
average	mix	mix	10.414	12.141	**15.585**
mix	average	average	10.418	**12.270**	15.322
mapping	average	average	10.376	12.137	15.467
masking	average	average	10.409	12.168	15.508
mix	masking	masking	**10.461**	12.252	15.491
mapping	masking	masking	10.397	12.134	15.539
average	masking	masking	10.390	12.159	15.466

在这个实验中，注意力集中在三个方面。每个头都是一个表示子空间[51]。图4-5显示了注意权重示例。从这个数字可以得出几个观察结果。总体而言，体重增加的注意力机制是单调进行的。在某种程度上，这表明本文使用的注意力机制有效[52]。详细地，每个框架及其相邻框架趋于具有更大的重量。我们使用类似于[53]中的注意力机制。

图4-6显示了频谱图的大小。所有增强的方法都会在混合信号中获得大部分语音信号。但是，他们提取的语音信号仍然包含伴奏信号的一部分。所有增强的方法都可以在混合信号中隐藏的低频下恢复频谱。但是，他们在恢复高频方面很差。 M-Mapping频谱图的高频部分仍然有很多噪声，而M-Masking频谱图的高频部分去除了太多的声音信息。 M-LSF通过平均M-Mapping和M-Masking做出

折衷，但是高频恢复仍然不够好。尽管MDM-tend-Separate-info（+ phase）方法在高频下仍然有一些噪音，但与其他方法相比有一些改进，特别是恢复了一些高频细节。这可能是因为在融合过程中，高频部分优先选择屏蔽频谱图，部分合并了映射频谱图的某些信息。

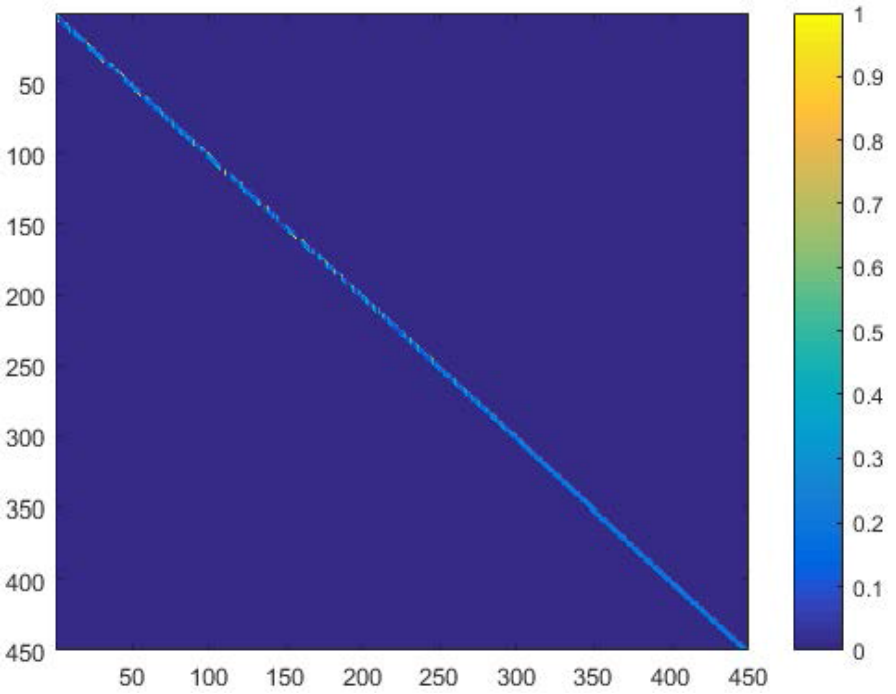
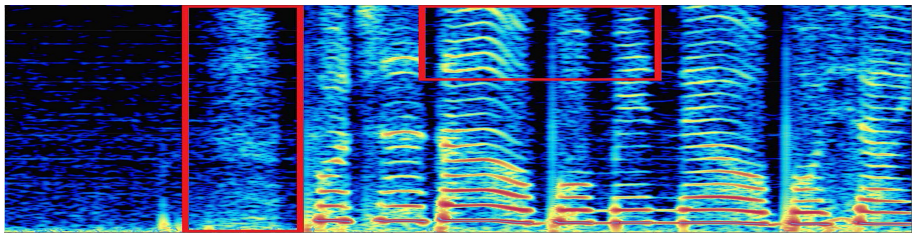


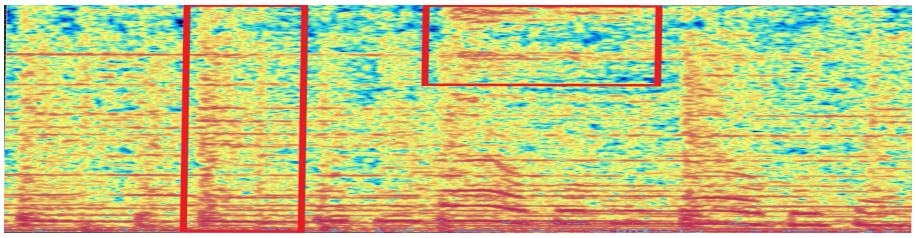
图 4-5　注意力权重示例：垂直轴索引和水平轴索引分别对应于语谱图中的帧，颜色越深表示值越大。

4.7　本章小结

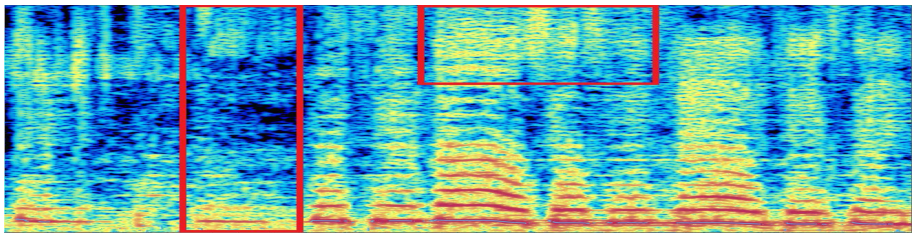
最小差异掩码（MDM）[50]显示出强大的增强能力，尤其是对于SIR和PESQ。在MIR-1K数据集上进行的实验表明，我们的系统一致且显着改善了定量评估。首先，常规术语可以帮助系统在SDR，SAR和SIR上获得更好的性能。其次，我们使用线性融合构造波形的相位来重构最终的增强波形，从而可以改善所有定量评估性能。此外，不同的嵌入方式提供不同的增强效果，并且我们观察到MDM-tend-Separate-info具有最佳的建模能力。在某种程度上，通过使用频谱图建模的注意力机制可以给出频谱图中相邻帧之间的相关性。注意机制为我们提供了一个新的想法，即在频谱图中找到关键帧可能有助于语音增强，这是我们未来的工作。



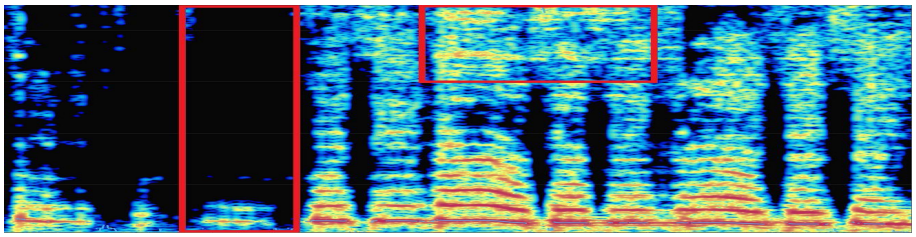
(a) Clean



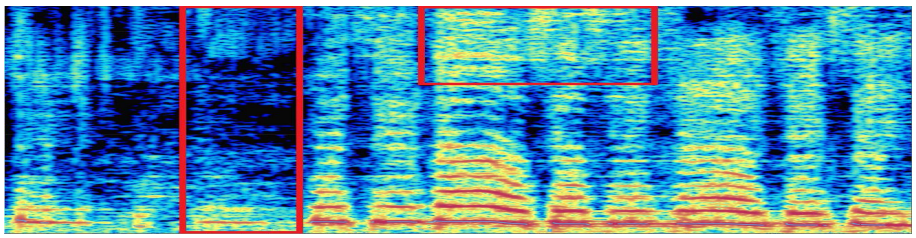
(b) Mix



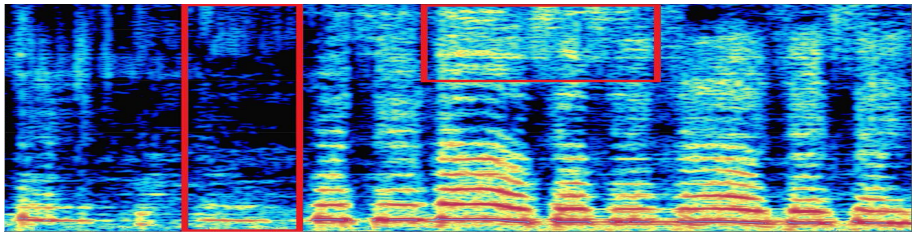
(c) M-Mapping



(d) M-Masking



(e) M-LSF



(f) MDM-tend-Separate-info

图 4-6 语谱图振幅信息：Clean表示干净语谱图的振幅信息，Mix表示带噪语谱图的振幅信息（未处理的语音信号），M-Mapping表示利用多目标学习得到的基于映射的增强语谱图振幅信息，M-Masking表示利用多目标学习得到的基于掩蔽的增强语谱图振幅信息，MDM-tend-Separate-info表示利用本节基于注意力机制的语谱图融合系统得到的增强语谱图振幅信息；横轴表示时间，纵轴表示频域，颜色越深表示值越大。

第 5 章 结语

5.1 总结

在这篇论文中，我们研究基于语谱图融合的语音增强方法。我们提出了基于最小差别掩蔽的语谱图融合系统，并在这个系统的基础上，设计了基于注意力机制的语谱图融合系统。我们分别在混响和加性噪声数据集下设计实验，并验证了本文提出方法的有效性。

针对语音混响，我们提出了一种带有最小差分掩膜估计系统的非线性频谱图融合方法。首先，进行了基于Bi-LSTM的多目标学习语音去混响前端，以获取不同的学习目标声谱图。然后，估计最小差别掩蔽将最接近标签的时频点分类。最后，我们使用第一阶段的频谱图和第二阶段的最小差别掩蔽来融合频谱图的最佳部分。当使用多目标学习时，我们观察到了频谱图和最小差别掩蔽（MDM）之间的主动特征互补。通过非线性频谱图融合，语音去混响主要提高了语音质量和语音混响调制能量比，例如，所有模拟数据的平均PESQ增益为0.1，而实际数据的平均SRMR增益为1.22。

针对加性噪声，我们提出了基于注意力机制的语谱图融合系统最小差异掩码（MDM）[50]显示出强大的增强能力，尤其是对于SIR和PESQ。在MIR-1K数据集上进行的实验表明，我们的系统一致且显着改善了定量评估。首先，常规术语可以帮助系统在SDR，SAR和SIR上获得更好的性能。其次，我们使用线性融合构造波形的相位来重构最终的增强波形，从而可以改善所有定量评估性能。此外，不同的嵌入方式提供不同的增强效果，并且我们观察到MDM-tend-Separate-info具有最佳的建模能力。在某种程度上，通过使用频谱图建模的注意力机制可以给出频谱图中相邻帧之间的相关性。

5.2 展望

在以后的研究中，我们将分析频谱图，并使用频谱图中的时变信息进行融合。此外，还将探索其他语音任务的特征融合，例如MFCC，以实现自动语音识别。

注意机制为我们提供了一个新的想法，即在频谱图中找到关键帧可能有助于语音增强，这是我们未来的工作。

此外，本文仅仅利用了直接映射和基于掩蔽的语音增强系统之间的互补性，但是仍然没有分析他们之间为什么会互补。在未来的工作中，我们会利用语谱图融合系统对他们之间的互补性进行验证。

参考文献

[1] Ortega-Garcia J, Gonzalez-Rodriguez J. Overview of speech enhancement techniques for automatic speaker recognition [C]. In Proceeding of International Conference on Spoken Language Processing, 1996: 929–932.

[2] Weninger F, Erdogan H, Watanabe S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR [C]. In Proc. International Conference on Latent Variable Analysis and Signal Separation, 2015: 91–99.

[3] Zhao X, Wang Y, Wang D. Robust Speaker Identification in Noisy and Reverberant Conditions [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22 (4): 836–845.

[4] Wang D, Chen J. Supervised Speech Separation Based on Deep Learning: An Overview [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26 (10): 1702–1726.

[5] Meyer J, Simmer K U. Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction [C]. In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997: 1167–1170.

[6] Ephraim Y, Van Trees H L. A signal subspace approach for speech enhancement [C]. In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993: 355–358.

[7] Varga A P, Moore R K. Hidden Markov model decomposition of speech and noise [C]. In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990: 845–848.

[8] Cao L, Zhang T, Gao H, et al. Multi-band spectral subtraction method combined with auditory masking properties for speech enhancement [C]. In Proceeding of International Congress on Image and Signal Processing, 2012: 72–76.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. nature, 2015, 521 (7553): 436–444.

[10] Xu Y, Du J, Huang Z, et al. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement [J]. arXiv preprint arXiv:1703.07172, 2017.

[11] Mallapragada P K, Jin R, Jain A K, et al. SemiBoost: Boosting for Semi-Supervised Learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31 (11): 2000–2014.

[12] Xu Y, Du J, Dai L, et al. An Experimental Study on Speech Enhancement Based on Deep Neural Networks [J]. IEEE Signal Processing Letters, 2014, 21 (1): 65–68.

[13] Sailor H B, Patil H A. Filterbank learning using Convolutional Restricted Boltzmann Machine for speech recognition [C]. In Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 5895–5899.

[14] Martin R. Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors [C]. In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002: 253–256.

[15] Xu Y, Du J, Dai L, et al. A Regression Approach to Speech Enhancement Based on Deep Neural Networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23 (1): 7–19.

[16] Donahue C, Li B, Prabhavalkar R. Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018: 5024–5028.

[17] Park S R, Lee J. A fully convolutional neural network for speech enhancement [J]. arXiv preprint arXiv:1609.07132, 2016.

[18] Sun L, Du J, Dai L, et al. Multiple-target deep learning for LSTM-RNN based speech enhancement [C]. In Procedding of Hands-free Speech Communications and Microphone Arrays, 2017: 136–140.

[19] Ge M, Wang L, Li N, et al. Environment-dependent attention-driven recurrent convolutional neural network for robust speech enhancement [C]. In Proceeding of Annual Conference of the International Speech Communication Association, 2019: 3153–3157.

[20] Choi H, Kim J, Huh J, et al. Phase-aware speech enhancement with deep complex U-Net [J]. arXiv preprint arXiv:1903.03107, 2019.

[21] Pascual S, Bonafonte A, Serrà J. SEGAN: Speech Enhancement Generative Adversarial Network [C]. In Proceeding of Annuual Conference of the International Speech Communication Association, 2017: 3642–3646.

[22] Gordon V S, Reda A. Trappy Minimax - using Iterative Deepening to Identify and Set Traps in Two-Player Games [C]. In Procedding of IEEE Symposium on Computational Intelligence and Games, 2006: 205–210.

[23] Wang Z, Bovik A C. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures [J]. IEEE Signal Processing Magazine, 2009, 26 (1): 98–117.

[24] Kolbæk M, Tan Z, Jensen J. Monaural Speech Enhancement Using Deep Neural Networks by Maximizing a Short-Time Objective Intelligibility Measure [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018: 5059–5063.

[25] Zhao Y, Xu B, Giri R, et al. Perceptually Guided Speech Enhancement Using Deep Neural Networks [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018: 5074–5078.

[26] Cohen I. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator [J]. IEEE Signal Processing Letters, 2002, 9 (4): 113–116.

[27] Fu S, Tsao Y, Lu X. SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement. [C]. In Proceeding of Annual Conference of the International Speech Communication Association, 2016: 3768–3772.

[28] Gao T, Du J, Dai L, et al. Densely Connected Progressive Learning for LSTM-Based Speech Enhancement [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018: 5054–5058.

[29] Wang Z, Wang D. Recurrent deep stacking networks for supervised speech separation [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2017: 71–75.

[30] Sun L, Li K, Wang H, et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training [C]. In Procedding of IEEE International Conference on Multimedia and Expo, 2016: 1–6.

[31] Du Z, Lei M, Han J, et al. Pan: Phoneme-Aware Network for Monaural Speech Enhancement [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 6634–6638.

[32] Xu Y, Du J, Dai L, et al. Cross-language transfer learning for deep neural network based speech enhancement [C]. In Procedding of International Symposium on Chinese Spoken Language Processing, 2014: 336–340.

[33] Pan S J, Yang Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22 (10): 1345–1359.

[34] Mathur A, Saxena V, Singh S K. Understanding sarcasm in speech using mel-frequency cepstral coefficent [C]. In Procedding of International Conference on Cloud Computing, Data Science Engineering - Confluence, 2017: 728–732.

[35] Chen Z, Watanabe S, Erdogan H, et al. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks [C]. In Proceeding of Annual Conference of the International Speech Communication Association, 2015: 3274–3278.

[36] Fu S, Wang T, Tsao Y, et al. End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26 (9): 1570–1584.

[37] Jung C, Joo Y, Kang H. Waveform Interpolation-Based Speech Analysis/Synthesis for HMM-Based TTS Systems [J]. IEEE Signal Processing Letters, 2012, 19 (12): 809–812.

[38] Fu S, Tsao Y, Lu X, et al. Raw waveform-based speech enhancement by fully convolutional networks [C]. In Procedding of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2017: 006–012.

[39] Wisdom S, Hershey J R, Wilson K, et al. Differentiable Consistency Constraints for Improved Deep Speech Enhancement [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2019: 900–904.

[40] Wang D, Jae Lim. The unimportance of phase in speech enhancement [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1982, 30 (4): 679–681.

[41] Han K, Wang Y, Wang D, et al. Learning Spectral Mapping for Speech Dereverberation and Denoising [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23 (6): 982–992.

[42] Griffin D, Jae Lim. Signal estimation from modified short-time Fourier transform [C]. In Procedding of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1983: 804–807.

[43] Greenberg S, Kingsbury B E D. The modulation spectrogram: in pursuit of an invariant representation of speech [C]. In Procedding of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997: 1647–1650.

[44] Paliwal K, Wójcicki K, Shannon B. The importance of phase in speech enhancement [J]. speech communication, 2011, 53 (4): 465–494.

[45] Weninger F, Hershey J R, Le Roux J, et al. Discriminatively trained recurrent neural networks for single-channel speech separation [C]. In Proceeding of GlobalSIP, 2014: 577–581.

[46] Rix A W, Beerends J G, Hollier M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs [C]. In Procedding of IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, 2001: 749–752.

[47] Kokkinakis K, Loizou P C. Evaluation of objective measures for quality assessment of reverberant speech [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2011: 2420–2423.

[48] Vincent E, Gribonval R, Fevotte C. Performance measurement in blind audio source separation [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14 (4): 1462–1469.

[49] Kolbæk M, Yu D, Tan Z, et al. Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25 (10): 1901–1913.

[50] Shi H, Wang L, Ge M, et al. Spectrograms Fusion with Minimum Difference Masks Estimation for Monaural Speech Dereverberation [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 7544–7548.

[51] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]. In Proceeding of neural information processing systems, 2017: 5998–6008.

[52] Raffel C, Luong M, Liu P J, et al. Online and linear-time attention by enforcing monotonic alignments [C]. In Proceeding of International Conference on Machine Learning, 2017: 2837–2846.

[53] Hao X, Shan C, Xu Y, et al. An Attention-based Neural Network Approach for Single Channel Speech Enhancement [C]. In Procedding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2019: 6895–6899.

发表论文和参加科研情况说明

（一）发表的学术论文

[1] **H. Shi**, L. Wang, M. Ge, S. Li, J. Dang, “Spectrograms Fusion with Minimum Difference Masks Estimation for Monaural Speech Dereverberation”, Proc. of ICASSP, 2020.

[2] **H. Shi**, L. Wang, S. Li, C. Ding, M. Ge, N. Li, J. Dang, H. Seki, “Singing Voice Extraction with Attention based Spectrograms Fusion”, Proc. of Interspeech, 2020.

[3] M. Ge, L. Wang, N. Li, **H. Shi**, J. Dang, X. Li, “Environment-dependent attention-driven recurrent convolutional neural network for robust speech enhancement”, Proc. of Interspeech, 2019.

[4] **H. Shi**, L. Wang, S. Li, M. Ge, N. Hou, E. S. Chng, J. Dang, T. Kawahara, “Mapping or Masking: Exploring Complementarities Between Them for Speech Enhancement”, (IEEE/ACM TASLP, under review).

[5] L. Qiang, **H. Shi**, L. Wang, S. Li, M. Ge, J. Dang, “Daub Learning: A Novel Progressive Learning Method for Speech Dereverberation” (ICASSP 2021, the joint first author, under review).

（二）参与的科研项目

[1] “基于语言认知机理的类脑自然语言识别与交互”，科技部国家重点研发计划“智能机器人”专项课题(No.2018YFB1305200)，2019.6-2022.5

[2] “面向机器人的复杂环境语音对话关键技术及系统实现”，新一代人工智能科技重大专项(No.18ZXZNGX00330)，2018.10-2021.9

[3] “面向混响环境的多口音语音识别研究”，国家自然科学基金面上项目(No.61771333)，2018.1-2021.12

致 谢

两年半的研究生生活马上就要过去了，很幸运能够进入天津大学智能与计算学部CCA组里，在这里留下了很美好的回忆!

感谢我的导师王龙标教授，在这两年半的时间里对我的悉心指导。感谢您给我机会让我融入到CCA这个大家庭，在这里，我学到很多。并且每次组会上，您都会认真解答我们的疑惑，并给我们想应的建议与意见，为我们的科研之路奠基。祝老师今后，工作顺利，身体健康!

感谢党建武教授在这两年半的时间里对我的指导。每次投会议之前的comments都十分必要而且很有用。并且您授课的激情和组会上open的指导对我们所有人都有很大的帮助。祝老师今后，工作顺利，身体健康!

感谢李胜老师对我全方位的指导！第一篇ICASSP是老师捞了我一手，不然肯定凉了。每次会议投稿前，您认真帮我看我的论文、找其他大佬帮我看论文，真的对我帮助很多很多！每次找您帮忙，您都会默默站在我们身后，有您真好！祝老师今后，工作顺利，身体健康!

感谢Chng教授对我的指导！给我期刊提供了一个很好的思路！和您交流真的很高兴，我特别喜欢和您有更多的交流，得到您更多的指点!

感谢葛檬师兄带我科研、带我各种玩。特别有幸能遇到这么好的一位师兄，把我从单打独斗中，拉到一个组共同奋斗。你带我读论文，想idea，而且把自己身边的人介绍给我！一句感谢真的太少太少！但很感谢你！祝师兄科研顺利，工作顺利，身体健康!

感谢崔凌赫师兄，我们一起打篮球，一起吃吃吃的时光真的很感动！很有幸在研究生阶段遇到你，从我入学，你带着我当了一年的实验室班长，学到了很多你做事的风格。篮球也是，不过你要多锻炼身体呀，再瘦二十斤！祝师兄工作顺利，家庭和睦，身体健康!

感谢司宇柯师姐，带我吃吃吃、转转转！尤其那次去Honda Seisei家里做饭，第一次做饭团，我自己都没想到我包的又大又圆…… 祝师姐科研顺利，家庭和睦，身体健康!

感谢郭丽丽师姐、刘佳星师兄、刘猛师兄、贡诚师兄、李楠师兄平时对我的帮助与指导！感谢!

感谢老铁宋世明，同样本科来自四川，话题很多，工作签的很顺利，以后不管在哪，有机会还是要一起吃饭，我看着你跑马拉松。

感谢我的舍友赵祥宇，不管是读博还是工作，希望大家以后还是多聚聚，天涯海角我有机会也会去找你玩！年轻就是好呀！

感谢林羽钦，目前同届唯一的一个博士！对我科研和生活上的帮助，也祝愿你在博士期间，成果多多，我们共同加油！

感谢郭少彤、杨婷、蔡心怡、周到、吴梦飞、吴双、徐杰、傅雅慧、张卓、陈帅婷在这两年多来的支持与鼓励，要是没有你们，我肯定不是现在的自己。

感谢师弟姜宇、强璐亚、尹浩然、陈森、秦思晴、汤丽、李盛兰、高嘉潞、王瑞芳、武艺博、齐剑书、吕永杰、黄武伟、刘大伟对我的帮助！

希望自己的变得最好，也祝愿实验室的小伙伴变得更好，更祝愿实验室能变得更好！